DOCUMENT RESUME

ED 222 038 FL 013 141

AUTHOR Svartvik, Jan; And Others
TITLE Survey of Spoken English: Report on Research 1975-81. Lund Studies in English, No. 63.
INSTITUTION Lund Univ. (Sweden). Dept. of English.
REPORT NO ISBN-91-40-04822-5
PUB DATE 82
NOTE 104p.

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS *Computational Linguistics; *Computer Programs; Databases; *English; Information Retrieval; Language Research; *Language Usage; *Oral Language; *Programing Languages

ABSTRACT

This is a report on the activities of the survey of spoken English at the University of Lund (Sweden), during the period 1975-81. The aim of the survey has been to make available in machine-readable form a corpus of material with its origin in speech. The corpus was built in conjunction with the survey of English usage project at the University of London. This report has six parts: (1) an introduction, (2) an account of the computerizing process, (3) a description of the output, (4) a summary of research projects in process, (5) lectures given at or arranged by the survey, and (6) a list of publications by the survey staff. The corpus is described as material consisting of 87 "texts," each containing 5000 words, with analysis of the basic prosodic distinctions of tone units, nuclei, boosters, and stresses. The process of editing, coding, transfer and storage, formatting and checking, correction, printing, and production of concordances is described. In a third part, the output in terms of texts, concordance, and word-lists is described and sources for printed versions and computer tape are given. In part 4, the review of research in progress includes illustrations of the ways the system is being used. The activities section provides names, addresses, and titles of lectures and seminars. Illustrations of aspects of the project are given in 15 figures and an appendix of word-class keys. (AMH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED222038

LUND STUDIES IN ENGLISH 63
EDITORS CLAES SCHAAR AND JAN SVARTVIK

SURVEY OF SPOKEN ENGLISH

Report on Research 1975–81

By
JAN SVARTVIK
MATS EEG-OLOFSSON
OSCAR FORSHEDEN
BENGT ORESTRÖM
CECILIA THAVENIUS

FL 013 141

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Jan Svartvik

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

CWK GLEERUP

LUND STUDIES IN ENGLISH
Editors: Claes Schaar and Jan Svartvik
Publishers: LiberLäromedel Lund

CWK Gleerup is the imprint for the scientific and scholarly publications of Liber Läromedel Lund.

Orders for publications in *Lund Studies in English* should be addressed to any international booksellers or directly to the publishers:
Liber
Box 1205
S-221 05 Lund, Sweden

This volume has been published with the aid of a grant from the Bank of Sweden Tercentenary Foundation.

© Survey of Spoken Enlish

ISBN 91-40-04822-5

Wallin & Dalholm Boktr AB, Lund 1982

Contents

Preface *9*

1 GENERAL *15*
- 1.1 Financial support *15*
- 1.2 Staff *15*
- 1.3 Premises *16*

2 COMPUTERIZING A CORPUS OF SPOKEN ENGLISH *17*
- 2.1 The Survey of English Usage/The Survey of Spoken English *17*
- 2.2 Editing of the texts *19*
- 2.3 Computer processing and output *25*
  - Coding *25*
  - Transfer and storage *25*
  - Formatting and checking *29*
  - Correction *29*
  - Printing *A Corpus of English Conversation* *30*
  - Production of concordances *30*

3 OUTPUT: TEXTS, CONCORDANCE AND WORD-LISTS *33*
- 3.1 Texts *33*
  - Description of tape and files *35*
  - Description of records *35*
- 3.2 Concordance *37*
  - Record description *37*
  - Sorting key *38*
  - Location in text *38*
  - Context *38*
  - Fields in the sorting key *39*
  - Text location reference fields *40*
  - Context field *40*
  - Sorting sequence *40*
  - Collating sequence *41*
- 3.3 Word-lists *41*

4 RESEARCH BASED ON THE LONDON-LUND CORPUS *51*
4.1 Word-class tagging *51*
Principles of coding *51*
First design of an interactive tagging program *53*
Final design *64*
Output from the program *65*
Input to the program *65*
The lexicon *66*
Suffix and context frame rules *67*
Program performance *69*
Tag concordance *69*
4.2 Syntactic tagging *70*
The unit of syntactic analysis *70*
Description of the phrase analysis program *73*
Tagging of clause elements *83*
4.3 Negation *86*
4.4 Turn-taking and interruption *87*
4.5 Questions and responses *88*
4.6 Third person reference *89*
4.7 Other studies *89*

5 OTHER ACTIVITIES *91*
5.1 Lectures, conferences and symposia *91*
5.2 Seminars and colloquia arranged by the Survey *93*
5.3 Visitors to the Survey *94*

6 PUBLICATIONS BY SURVEY STAFF 1975-1982 *97*

REFERENCES *101*

APPENDIX: WORD-CLASS TAGS *105*

LIST OF TABLES, FIGURES AND APPENDIX

Tables

Table 1 The corpus of the Survey of English Usage *18*
Table 2 Text classification of the London-Lund Corpus *20-21*
Table 3 List of symbols in book, slip and tape versions *26-28*
Table 4 The 100 most frequent words in the London-Lund Corpus *43-46*
Table 5 The 100 most frequent words in the London-Lund Corpus compared with LOB and Brown *47-49*
Table 6 Distribution of prosody and grammatical functions of *that* *68*

Figures

Figure 1 The original paper slip *22*
Figure 2 Printout from the computer tape *23*
Figure 3 Printed version using a photosetting technique *24*
Figure 4 Printed version of Text S.1.1 *54-55*
Figure 5 Printout of a tagged text *56-62*
Figure 6 Interactive tagging system design *63*
Figure 7 Concordance sorted alphabetically by key-word in context *71*
Figure 8 Flowchart for verb phrase analysis *76-81*
Figure 9 Output from the verb phrase tagging program *84-85*

Appendix

Word-class tags *105*

Preface

AIMS OF THE SURVEY OF SPOKEN ENGLISH

This is a report on the activities at the Survey of Spoken English, University of Lund, during the period 1975-1981 which has been financed chiefly by the Bank of Sweden Tercentenary Foundation.

The Survey of Spoken English is a sister project of the Survey of English Usage at University College London. Its primary aim was to make available, in machine-readable form, the spoken material which has been collected by the Survey of English Usage. This task may seem a trivial one to those who have had no experience of working with large amounts of data from spoken language; those who have been actively engaged in such operations will testify to the time-consuming work and the numerous linguistic and technical problems that need to be solved in the process. An account of the relation between the two Surveys and of the computerizing of the corpus is given in section 2 of this report.

It is gratifying to be able to state, at the end of this six-year period, that the primary aim has been achieved. The text of the London-Lund Corpus of Spoken English is now available, in its entirety, on magnetic tape for computer use and, in part, in printed form. There is also a concordance available on computer tape, which should provide a research tool with many possible uses. A detailed specification of the output in the form of texts, concordance and word-lists is given in section 3 of this report.

In the application for funding, a secondary aim of the Survey of Spoken English was stated to be the use of the corpus for linguistic research. Unlike the primary task of producing a corpus, this is not a finite task. It is therefore natural that the research based on the corpus and presented in section 4, should be chiefly research in progress.

THE INTERNATIONAL COMPUTER ARCHIVE OF MODERN ENGLISH TEXTS

The first standardized, edited and generally available computerized corpus of English was the Brown University Standard Corpus of Present-Day American English, commonly known as *the Brown Corpus*. Produced under the direction of W. Nelson Francis it has now been in use all over the world for more than fifteen years (see eg Francis 1979; 1980; Francis & Kučera 1964; at press; Kučera & Francis 1967).

Recently, the joint efforts of Geoffrey Leech at the University of Lancaster and Stig Johansson at the University of Oslo in cooperation with the Norwegian Computing Centre for the Humanities (NAVF) at the University of Bergen have produced *the Lancaster-Oslo/Bergen Corpus*, a corpus of British English to match the Brown Corpus of American English (see Johansson 1982; Johansson, Leech & Goodluck 1978; Hofland & Johansson 1982).

With the advent of *the London-Lund Corpus* we now also have a collection of spoken British English material (see Svartvik & Quirk 1980 and other references in section 6).

All three corpora are distributed through the International Computer Archive of Modern English (ICAME) which was established in 1977 by a small group of scholars from England, the United States, Sweden, and Norway. The primary purposes of the organization are:

a) collecting and distributing information on (i) English language material available for computer processing; and (ii) linguistic research completed or in progress on this material;

b) compiling an archive of corpora to be located at the University of Bergen, from where copies of the material can be obtained at cost.

The Norwegian Computing Centre for the Humanities at Bergen serves as a clearing-house for ICAME, organizing the distribution of material and publishing the newsletter *ICAME News* with Stig Johansson, University of Oslo, as editor.

WHY A CORPUS OF SPOKEN ENGLISH?

It is now widely recognized that collections of material, so-called 'corpora', are useful, or even necessary, for some forms of linguistic research. There are many reasons for using a corpus, alone or in combination with elicitation techniques and introspection. A corpus makes it possible

- o to state a wide repertoire of uses to which a language is put;
- o to make more objective statements than introspective analysis permits;
- o to achieve total accountability of linguistic features;
- o to state frequencies of occurrence in different uses of the language;
- o to share with other researchers a common basis for analyses of similar or related phenomena;
- o to give researchers in any part of the world access to material that would otherwise be difficult or impossible to obtain.

For spoken English, the last point deserves to be singled out for special mention. Making recordings of spoken English may not at first sight seem to be a notable achievement any longer; but, when it comes to making, say, surreptitious recordings of informal conversations, a large number of the students of English in the world will be hard put to do it - we should not forget that the majority of these students do not in fact live in an environment where English is spoken as a native language. This corpus of spoken English has, furthermore, been transcribed and analyzed prosodically. This is a task which requires special skill, a skill that not all prospective users of this corpus will possess, or care to acquire, since their interest in the uses of this material is likely to be in an area of research quite different from that of prosody.

RESULTS

Since this corpus has only recently become generally available, the important results of research based on it are still to come. Most of the studies described in later sections of this report will be more adequately documented in future publications. At Lund and Uppsala, further work on all three corpora will be carried out in the project English in Speech and Writing which is supported by the Swedish Council for Research in the Humanities and Social Sciences. Instead of commenting on this project and other studies in progress, it might be more appropriate to say something in general about future uses of this corpus of spoken English.

In the course of this project the climate of linguistic research has undergone a change in the direction of a more positive attitude among linguists to the value of a corpus. There is now an increasing demand from various branches of science for material of this kind for the study of, for example, discourse, intonation (and other prosodic features), face-to-face interaction, language learning, speech processing, computer recognition of natural language, and so forth. There are numerous uses for a generally available sample of authentic and prosodically analysed spoken English such as the London-Lund Corpus.

ACKNOWLEDGEMENTS

It goes without saying that a project of this size and nature could not have been successfully completed without generous financial support from the Bank of Sweden Tercentenary Foundation, the help of a group of assistants devoted to the task, and good advice from colleagues abroad and at home. We have tried to list all these supporters in the next section of this report.

But in addition to support one needs inspiration and I would like to conclude this preface by mentioning two colleagues who have supplied both in ample measure. One is Randolph Quirk,

with whom I had the privilege of being associated at University College London during the early sixties and later as a corresponding co-author; the other is W Nelson Francis, with whom I have had several contacts over the last twenty years, most intensively during my period at Brown University. Both these scholars provided the inspiration for and belief in corpus study without which the Survey of Spoken English would not have seen the light of day.

Lund, April 1982
Jan Svartvik

1 General

1.1 FINANCIAL SUPPORT

The Survey of Spoken English has received its chief financial support from the Bank of Sweden Tercentenary Foundation during the whole period from 1975 to 1981, covering costs for staff, printing, books, equipment and travel expenses. The Knut and Alice Wallenberg Foundation has provided funds for the purchase of a video display terminal with cassette recorder. The University of Lund has provided premises and various foundations of the University have given travel grants to Survey staff. All grants have been administered by the University.

1.2 STAFF

The research group has consisted of the following persons:

Jan Svartvik, PhD, Professor, Director of the Survey
Mats Eeg-Olofsson, FK, Programmer (since 1979)
Oscar Forsheden, FL, Lektor (since 1977)
Penny Kenton Russ, Secretary (1976-1978)
Bengt Oreström, FK, Research assistant
Cecilia Thavenius, FM, Research assistant

Throughout the period, Bengt Oreström and Cecilia Thavenius have been in charge of the editing procedure and the daily running of the project. Mats Eeg-Olofsson has been responsible for all computer programming during the last two years. Penny Kenton Russ carried out the input of the material from a terminal in addition to doing secretarial work. Oscar Forsheden has taken on a number of duties, in particular proofreading.

All members of the research group have played an active part in the production of word-lists, concordances, and a

printed version of part of the corpus and also participated in the preparation of the tagging system.

The staff of the Computing Centre of Lund University, in particular Lennart Månsby and Sven Norén, did the programming during the first four years.

Two postgraduate students, Hans Lindqvist and Wiveca Sotto, were temporarily employed during the summer of 1980 to assist with word-class tagging.

Several members of the Survey of English Usage, University College London, have given us valuable assistance, in particular Noel Burton Roberts, René Quinault, Catherine Sweetman, and Janet Whitcut.

We have also benefitted from discussions with a number of colleagues in Sweden, in particular Sture Allén and Rolf Gavare, Göteborg; Benny Brodda and Jan Hultgren, Stockholm; Anna-Lena Sågvall Hein, Uppsala; Bengt Altenberg, Anna-Brita Stenström, Gunnel Tottie, and Sidney Wood, Lund.

Finally, for the production of this report, we want to express our thanks to Muriel Spalding-Larsson for reviewing our written English, to Andrzej Szlagor for drawing our figures and to Ann-Mari Gayle for typing our manuscript.

1.3 PREMISES

During most of our six-year period, the Survey has been housed in the old Porter's Lodge in the 'Paradis' area of the University. The premises have been adequate, consisting of three rooms, the largest of which has also been used for seminars and conferences.

2 Computerizing a Corpus of Spoken English

2.1 THE SURVEY OF ENGLISH USAGE/THE SURVEY OF SPOKEN ENGLISH

The aim of the Survey of English Usage, University College London, is to describe the grammatical repertoire of adult educated native speakers of British English. An important part of this task was the compilation of a corpus of material, comprising both written and spoken language. The design of the corpus, shown in Table 1 (p 18), is described in some detail in Svartvik & Quirk 1980.

All texts with origin in speech have been subjected to a detailed prosodic and paralinguistic analysis according to the system presented in Crystal & Quirk 1964 and Crystal 1969. The corpus is stored on paper slips in filing cabinets at University College London.

The Survey of Spoken English, University of Lund, began in 1975 with the primary aim of making the material with origin in speech of the Survey of English Usage available in machine-readable form. This would make it possible for interested scholars, in any part of the world, to gain access to large amounts of spoken English in orthographic representation with prosodic analysis and also to use a computer for machine processing.

The computerized version of the spoken English material will be referred to as the London-Lund Corpus of Spoken English, abbreviated LLC. The classification of the Corpus texts is shown in Table 2 (pp 20-21).

TABLE 1. THE CORPUS OF THE SURVEY OF ENGLISH USAGE

(I) Material with origin in writing (100 texts)

(A) Printed (46)

Learned arts		6
Learned sciences		7
Instructional		6
Press	general news	4
	specific reporting	4
Administrative & official material		4
Legal and statutory material		3
Persuasive writing		5
Prose fiction		7

(B) Non-printed (36)

Continuous writing	imaginative	5
	informative	6
Letters: social	intimate	6
	equal	3
	distant	4
Letters: non-social	equal	4
	distant	4
Personal journals (diaries)		4

(C) As Spoken (18)

Drama		4
Formal scripted oration		3
Broadcast news		3
Talks	informative	4
	imaginative	2
Stories		2

(II) Material with origin in speech (100 texts)

(A) Monologue (24)

Prepared (but unscripted) oration			6
Spontaneous	oration		10
	commentary	sport	4
		non-sport	4

(B) Dialogue (76)

Conversation	Surreptitious	intimate	24
		distant	10
	Non-surreptitious	intimate	20
		distant	6
	Telephone	intimate	10
		distant	6

2.2 *EDITING OF THE TEXTS*

The transfer of the corpus to computer tape was preceded by a long period of editing and checking of the available corpus, which consists of 87 'texts', each of 5000 words. After many discussions we decided to reduce the number of features in the original analysis of the corpus for our future computational work. Thus, the basic prosodic distinctions of tone units, nuclei, boosters, and stresses have been retained, whereas other features have been omitted, including tempo (allegro, clipped, drawl, etc), loudness (piano, forte, etc), modifications in voice quality (pitch, range, rhythmicality, and tension), voice qualifiers (whisper, creak, etc), and voice qualifications (laugh, sob, etc).

The reasons for reducing the number of features were partly practical and technical, partly linguistic. While we do not want to minimize the importance of paralinguistic features, it is clear that they are less central than the basic distinctions (such as size of tone unit, tone, place of nucleus) for most grammatical studies of spoken English. However, for those users who require the full set of features, the original material on paper slips will of course continue to be available for reference at the Survey of English Usage in London.

To illustrate the different representations of the material, three samples are given in Figures 1, 2 and 3. Figure 1 shows the original representation of the material on a paper slip; Figure 2 shows the same text sample on computer printout with the reduced set of features; and Figure 3 shows the text in its printed version (as in Svartvik & Quirk 1980).

The editorial task, which was carried out by the assistants, required great accuracy and turned out to be a fairly slow process. As a result of this meticulous work where each text was scrutinized by at least three people, a number of inconsistencies in the original London version were discovered, but they could be corrected through correspondence with and visits to the Survey of English Usage in London.

TABLE 2. TEXT CLASSIFICATION OF THE LONDON-LUND CORPUS

TEXT GROUPS		Subgroups	Dialogue	Face-to-face	Private	Surreptitious	Radio	TEXT LABELS	SUB-GROUP TOTAL	TEXT GROUP TOTAL
ABBREVIATION	BROAD DESCRIPTION									
CON	Face-to-face conversation	A	+	+	+	+	-	S.1.1-14, S.2.1-14, S.3.1-6	34	46
		B	+	+	+	-	-	S.4.1-7, S.5.8-11, S.6.2 [1]	12	
TEL	Telephone conversation	C	+	-	+	+	-	S.7.1-3, S.8.1-4, S.9.1-3 [2]	10	10
DIS	Discussion, interview, debate	D	+	+	-	-	+	S.5.1-7, S.6.1, S.6.3-5 [3], S.10.8 [4]	12	12
PUB	Public, unprepared commentary, demonstration, oration	E	+	+	-	-	-	S.11.1, S.11.4-5	3	12
		F	-	+	-	-	+	S.6.6 [5], S.11.3	2	
		G	-	-	-	-	+	S.10.1-7 [6]	7	
PRE	Public, prepared oration	H	-	+	-	-	-	S.11.2, S.12.1-6 [7]	7	7
			Number of "+" texts: 71	70	56	44	21		87	87

NOTES TO TABLE 2

1 Subgroup B is composite; individual speakers can be marked for surreptitiousness.

2 S.9.3 is ansaphone calls.

3 Subtext S.6.4b, which constitutes only a minor part of Text S.6.4, is not a radio recording but a discussion between an ex-nurse and a research worker on a topic similar to that of S.6.4a.

4 Subtext S.10.8b is not a radio recording. It has been placed in S.10.8 because it is very short, less than 100 tone units. Furthermore, the speaker in S.10.8b is one of the speakers in S.10.8a and both subtexts deal with popular science.

5 S.6.6 is an interview which, as it turned out, virtually became a monologue.

6 Subtext S.10.6a is a TV recording. Subtext S.10.7c is not a radio recording (like the main part of Text S.10.7) but a demonstration.

7 There are two speakers in S.12.3. However, it has been placed in this text group since it is virtually a monologue.

FIGURE 1. THE ORIGINAL PAPER SLIP

S.1.3-9
•bro•chùre• for# so I g /díd it#g# . and /then ¯ anóther
•one#p# - and
B • mhm •
(A) /Nthèn they •said# well "/now that you've done Nthése#
and they've been a "/sò sucNcéssfu_#a# we'd /like you
to do our l Nsùpēr# . /alpha:màtic#l# a or /Nsómething#
m:narrow and /this is one of Nthése#a# that /goes mm' sídeways#
m':rhythmic and /frǿntwards# and em/bróíders# and •/dárns#m'# and
sews •/bùttons on#m#
B •(- laughs) yes•
(A) -- a and I /sīid# well a# I /don't Nrèally •think# I
m:slurred could /wríte# -- a and this was a m sort of m#a#
/ninety six page :bòoklet# p /you Nknów# p# about /that
Ntìg# •-• em I'd I'd /used to gò through# /each of the
B • m •
(A) •processes at :hòme# •.• a I ¯don't think it will be
•e/nough a# just to have

FIGURE 2. PRINTOUT FROM THE COMPUTER TAPE

```
1 3    9 1410 1 1(A     11   ‖th↘en they ▷said■
1 3    9 1430 1 1 A     11   well .‖now that you've done th↗ese■
1 3    9 1440 1 1 A     11   and they've been .‖s↘o succ↗essful■
1 3    9 1450 1 1 A     11   we'd ‖like you to do our s↘uper■ ·
1 3    9 1460 1 1 A     11   ‖alpha:m↘atic■
1 3    9 1470 1 1 A     11   or ‖s↗omething■
1 3    9 1480 1 1 A     11   and ‖this is one of th↗ese■
1 3    9 1490 1 1 A     11   that ‖goes s↗ideways■
1 3    9 1500 1 1 A     11   and ‖fr↗ontwards■
1 3    9 1510 1 1 A     11   and em‖br↗oiders■
1 3    9 1520 1 1 A     11   and ★‖d↗arns■
1 3    9 1530 1 1 A     11   and sews★ ‖b↘uttons on■
1 3    9 1540 1 1 b     20   ★( - laughs) yes★
1 3    9 1550 1 1(A     11   - - and I ‖s→aid■
1 3    9 1560 1 1 A     11   well I ‖don't r↘eally ▷think■
1 3    9 1570 1 1 A     11   I could ‖wr↘ite■ - -
1 3    9 1580 1 1 A     11   and this was a sort of ‖ninety-six page :b↘ooklet■
1 3    9 1590 1 1 A     11   ‖you kn↗ow■
1 3    9 1600 1 1 A     11   about ‖that b↘ig■ ★-★
1 3    9 1610 1 1 A     11   [ə:m] I'd I'd ‖need to g↘o through■
1 3    9 1620 2 1 A     21   ‖each of the
1 3    9 1630 1 1 b     20   ★[m]★
1 3   10 1620 1 1(A     11   processes at :h↘ome■ ★·★
```

FIGURE 3. PRINTED VERSION USING A PHOTOSETTING TECHNIQUE

> A 141 ‖THÈN they ▷said■ 143 well ˈnow that you've done THÉSE■ 144 and they've been ˈSÒ SUCCÉSSFUL■ 145 we'd ‖like you to do our SÙPER■ · 146 ‖ALPHA▵MÀTIC■ 147 or ‖SÓMETHING■ 148 and ‖this is one of THÉSE■ 149 that ‖goes SÍDEWAYS■ 150 and ‖FRÓNTWARDS■ 151 and EM‖BRÓIDERS■ 152 and ☆‖DÁRNS■ 153 and sews☆ ‖BÙTTONS on■

b 154 ☆(- laughs) yes☆

> A 155 - - and I ‖SÂID■ 156 well I ‖don't RÈALLY ▷think■ 157 I could ‖WRÌTE■ - - 158 and this was a sort of ‖ninety-six page ▵BÒOKLET■ 159 ‖you KNÓW■ 160 about ‖that BÌG■ ☆-☆ 161 [əm] I'd I'd ‖need to GÒ thr.ougʌ■ 162 ‖each of the

b 163 ☆[m]☆

> A 162 processes at ▵HÒME■ ☆ · ☆

2.3 COMPUTER PROCESSING AND OUTPUT

The computer processing in the first phase of the project included the input, storing, checking, correction, and printout of the corpus. Sven Norén and Lennart Månsby have been responsible for this programming. They have also produced a first version of the concordance, which has later been modified by Mats Eeg-Olofsson who has also designed and implemented the system for computer-aided tagging of the corpus undertaken during the second phase of the project. The programming for phototype-setting the printed part of the corpus from magnetic tape has been done by Lars Aspsäter, LiberTryck.

Coding

The information selected from the original Survey of English Usage slips was typed in off-line on a Delta 5000 video display terminal and stored on cassettes in a Datum model 4200 digital cassette tape transport connected to the terminal via a Datum model 4000 terminal adapter. During this input operation certain special signs (phonetic, prosodic, etc) were manually translated into sequences of ASCII characters. (See Table 3 for the correspondence between the sign representation in the different versions - book, slip, tape - of the corpus.)

Transfer and storage

The texts were transferred from the cassettes to a mass storage file in the UNIVAC 1108 computer at Lund University Computing Centre via an on-line modem connection. Technically, this was accomplished by using the cassette tape transport in the same way as a paper tape reader connected to a teletype terminal.

TABLE 3. LIST OF SYMBOLS IN BOOK, SLIP AND TAPE VERSIONS (From Svartvik & Quirk 1980)

The following section lists the symbols used in the transcription. The material now exists in three different versions: on paper slips in the Survey of English Usage, University College London and the Survey of Spoken English, University of Lund (SLIP); on magnetic computer tape (TAPE); and in the printed form of this book (BOOK). Since some users of this corpus may want to use or refer to more than one version, the symbols used in the three versions have been listed below, side by side. It is not possible to explain the symbols here and for explanations of the prosodic system we refer to Quirk et al 1972 (Appendix II), Crystal 1969 and Crystal and Davy 1975.

	BOOK	SLIP	TAPE	
TEXT NUMBER	S.2.3 S.3.1a	S.2.3 S.3.1a		Text number; small letter denotes subtext (example)[1]
SPEAKER	A	A		Speaker identity (example)[2]
	>A	(A)		Speaker identity; speaker continues where he left off (example)
	A, B	A, B		A and B (example)[3]
	A/B	A/B		A or B (example)[4]
	VAR	various		Various speakers
	?	?		Speaker identity unknown
	▪	NSA		Non-surreptitious speaker (example)[5]
	☆yes☆ +yes+	*yes* +yes+	042 | 043 |	Simultaneous talk[6]
	(laughs)	(laughs)	040 041	Contextual comment[7]
	«yes»	((yes))	040 040 041 041	Incomprehensible words[8]
TONE UNIT	■	•	035	End of tone unit (TU)[9]
	‖	/	094	Onset
	[yes]	[yes]	123 125	Subordinate TU[10]
NUCLEUS[11]	yès	yès	092	Fall
	yés	yés	047	Rise
	yēs	yēs	061	Level
	yès yés	yès yēs	092 047	(Rise-) fall-rise[12]
	yès yés	yès yēs	047 092	(Fall-) rise-fall
	yès yés	yès yès	092 047	Fall-plus-rise
	yés yès	yès yès	047 092	Rise-plus-fall

	BOOK	SLIP	TAPE	
BOOSTER[13]	▷yes	ˈyes	095	Continuance
	▵yes	ːyes	058	Higher than preceding syllable[14]
	△yes	ˈyes	033	Higher than preceding pitch-prominent syllable
	Δyes	ˈˈyes	033 033	Very high
STRESS[11]	ˈyes	ˈyes	039	Normal
	ˈyes	ˈyes	034	Heavy
PAUSE[16]	yes · yes	yes · yes	046	Brief pause (of one light syllable)
	yes — yes	yes — yes	045	Unit pause (of one stress unit or 'foot')
	· ·	· ·	045 046	Combinations of pause
	— —	— —	045 045	
	— — ·	— — ·	045 045 046	
	— — —	— — —	045 045 045	
GRAPHICS	full-time	full-time	045	Hyphen[17]
	that's	that's	096	Apostrophe
	à	à	097 052	eg *à la maison*
	ä	ä	097 053	eg *Händel*
	é	é	101 049	eg *donné*
	è	è	101 050	eg *très*
	ê	ê	101 051	eg *être*
	ë	ë	101 052	eg *Citroën*
	ö	ö	111 049	eg *Höllgarten*
	ü	ü	117 049	eg *Dürer*
	ç	ç	099 053	eg *français*
PHONETICS[18]	[ə]	[ə]	064	
	[tʃ]	[tʃ]	116 115 104	
	[dʒ]	[dʒ]	100 122 104	
	[θ]	[θ]	116 104	
	[ð]	[ð]	100 104	
	[ʃ]	[ʃ]	115 104	
	[ʒ]	[ʒ]	122 104	
	[ŋ]	[ŋ]	110 103	
	[ɑ]	[ɑ], [ı]	126	Voiceless clicks
	[œ]	[oe]	111 101	eg in Fr. *peur*
	[ʔ]	[ʔ]	063	Glottal stop
	[ã]	[ã]	097 049	Nasal

NOTES

1 Text number on the magnetic computer tape can be found in columns 1 – 5.

2 Speaker identity on the magnetic computer tape is given in columns 18 – 23. Note that '?' (063), '/' (047), and '(' (040) have different functions in columns 18 – 23 and columns 28 – 78.

3 A and B speak, laugh, etc. simultaneously.

4 Either A or B speaks, laughs, etc.; speaker identity cannot be firmly established.

5 For the term 'non-surreptitious speaker'.

6 Simultaneous talk is indicated by pairs of asterisks or plus signs, and it occurs in various combinations:

A ✱ . . . ✱
B ✱ . . . ✱

A ✱ . . .
B ✱ . . . ✱
C . . .
D . . . ✱

} Everything from the first to the last asterisk occurs simultaneously, i.e. also C's talk.

A ✱ . . .
B ✱ . . . ✱
A . . . ✱

A . . ✱ . . . ✱ . .
B ✱ . . . + ✱ . . . +

A ✱ . . .
B ✱ . . . ✱
C . . . ✱

C + ... +

7 'Contextual comment' indicates non-linguistic activity, such as ‹laughs›, (coughs), (enters), (telephone rings) or technical mishaps, e.g. (gap in recording), (10 seconds untranscribable). Note that '‹' and '(' here mean length of contextual comment, e.g. (‹ laughs, (‹ coughs), (‹ ‹ giggles), etc.

8 'Incomprehensible words' denotes one or more words where it is uncertain what is said in the tape-recording. In some cases it is only possible to tell the number of syllables, e. g. «10 to 12 syll›».

9 The 'tone unit' (TU) is the basic prosodic unit in our analysis. The 'onset' is the first prominent syllable in a tone unit. It is possible for a speaker to complete another speaker's TU: e. g.

B ‖ and they sort of REINFORCE one ‹an‖other's [ɑ:] . «sort of»
D ΔEGOS ◉‖◉ (S.1.12 TU 989-990)

10 Subordinate TUs can occur in complex combinations, e. g.

well ‖why 'should ▷ they ▷ not ▷ use ‹the ‖words of the ORIGINAL ‹if they're ‖making a SUMMARY◉‖ ◉‖ ◉ (S.1.1 TU 964)

11 The peak of greatest prominence in a tone unit is called 'nucleus'. The nucleus symbol has been placed as follows. In words with a single vowel: YĒS, YĒS, YĒS, in digraphs: RĒAD, in words with a mute vowel: GUESS, in words with a semivowel: BĒAUTIFUL, EUROPE but QUITE, CHOIR. A consonant can also take nucleus, e. g. OK, [m̩]. Throughout this book the word carrying the nuclear tone is printed in small capital letters except for 'words' appearing within square brackets. On the magnetic tape the nucleus symbol is placed before vowel/consonant.

12 In addition to the combinations of nucleus types listed here, there are other combinations, such as fall-rise plus fall, fall plus rise plus fall, fall plus level etc. (see Crystal 1969). The latter part of trinuclear tones determines with what binuclear tones they are classified.

13 Whereas direction of pitch is described in terms of nucleus, range of pitch is described by the booster system, which indicates relative pitch level. For stressed syllables the normal situation is a slight drop. Departures from this have been marked with the given symbols placed before the syllable boundary.

14 Colon (058 in decimal code) has two functions on the magnetic tape. Outside square brackets it always denotes booster; inside square brackets it denotes vowel length in all cases except when it is immediately preceded by an opening square bracket '[' or space, in which case it denotes booster, e. g.

:yes = booster [:ði] = booster
[ði:] = vowel length [ði :ði] = booster

15 Stress symbols are placed before syllable boundary.

16 Pause symbols are always surrounded by space except in combination with asterisks '✱' and plus signs '+', both denoting simultaneous speaker activity.

17 Hyphen as punctuation mark is never preceded or followed by space except when pause is included in a hyphenated word. Compare note 7 and the following:

full-time = hyphenated word *(full-time)*
full - time = *full* + pause + *time*
full- - time = *full* + hyphen + pause + *time*

A fourth use of the hyphen on the magnetic computer tape is as a connecting symbol for words which are normally written solid but are split up in the text by a pause, e. g.

remem . ber

On the magnetic computer tape such word partials surrounding a pause are indicated by a hyphen placed immediately in front of each of the two word partials, e. g.

-remem . -ber

In this case the hyphen is thus used as a technical device, not as a true graphic symbol.

18 All phonetic symbols are given in square brackets (in decimal tape code '[' = 091, ']' = 093).

The whole corpus was stored on disc as a 'program file', in the jargon of the UNIVAC operating system EXEC. The texts were stored sequentially as 'symbolic elements' of the program file. This enabled us to use standard editing software for a large part of the processing. For security reasons, the corpus file was frequently copied onto magnetic tape. Besides acting as a back-up, mag tape storage was also used for transferring preliminary versions of the corpus to other computers.

Formatting and checking

On transfer to disc the texts were divided into tone units, which were provided with numbers, and further split up into links and records. (See p 35 for a detailed description of the records.) After being formatted, the texts were checked by a program spotting, among other things, illegal or suspect combinations of prosodic signs, and making sure that the different kinds of brackets in the texts were properly nested and matched.

Correction

After input, formatting and checking, the texts were printed on paper by a line-printer, and proofread by the Survey staff. Observations made in the proofreading as well as in the automatic checking of the texts were coded and stored on cassettes. These corrections were transferred to the computer in the same way as the texts, and used as commands to the standard editor ED, which was applied to the texts.

After correction the texts were again checked automatically and printed. This correction cycle was repeated several times. For some last-stage corrections we used the interactive editing facilities of the UNIVAC time-sharing system CTS.

Proofreading was in fact carried out six times for each of the 87 texts engaging all the assistants. Moreover, check programs were run after the fourth stage of proofreading to spot logical errors that had not yet been found through the manual checking. This whole procedure turned out to be extremely time-consuming and in fact much more laborious than any of us had been able to foresee.

Complications with the orthographic representation also had to be faced and taken care of. Written variants (eg *-ise/ize*, *Royal/royal*, *flower pot/flower-pot/flowerpot*), which are of importance for the compilation of word lists, etc, were standardized. However, as there is no full consistency in the language in this respect, we had to accept a certain degree of inconsistency also in our material.

Printing "A Corpus of English Conversation"

Since 34 texts representing spontaneous conversation were to be published in printed form, a great deal of discussion was devoted to finding ways of improving on the representation of prosodic symbols on the original typewritten slips as well as the symbols on the computational printouts. LiberTryck, Stockholm, handled the advanced programming procedure of transferring the symbols to the printed version which was produced by a photosetting technique from the magnetic computer tape. Examples of the printed version are reproduced on pp 54-55.

Production of concordances

Concordances of the texts were produced to facilitate proofreading as well as for the purposes of further research. Since no general concordance-making program was available at the Computing Centre, and it seemed unlikely that a general program could cope with the complicated structure of the texts in the corpus, we designed a special-purpose program (see Norén 1979).

It took about one man-month's work to implement the program, which was written in FORTRAN. The size of the source code is approximately 1000 lines. The program proved quite efficient. Its modular design has facilitated maintenance and revision.

The basic operation of the concordance program is to create, for every instance of a key word, a record consisting of the key word in its context and a detailed sorting key. The records are then sorted by the standard sort/merge package, stored on tape, and printed. The concordance can be printed selectively by means of a special program which extracts only those instances that match certain simple graphic patterns, ie contain certain letter sequences input as parameters to the program.

The concordance has also been used for making frequency word lists.

3 Output: Texts, Concordance and Word-Lists

The computer processing of the corpus has produced texts, concordances and word-lists.

3.1 TEXTS

All the texts are available on magnetic computer tape. (See example of output on p 23.) On the tapes the text corpus proper is preceded by a detailed account of the texts and the speakers. A classification of the whole corpus is reproduced on p 20 in terms of the features dialogue, face-to-face, private, surreptitious and radio.

Subgroup A in the text classification, consisting of 34 conversation texts, has also been printed in a book, *A corpus of English conversation*, ed by J Svartvik and R Quirk (1980).

THE BOOK

The printed version, *A corpus of English conversation*, which also includes a presentation of the two surveys, a complete description of the corpus, a list of symbols, information about the speakers, and a description of the computer tape version, has been published in the series *Lund Studies in English* (893 pp) by

CWK Gleerup Publishers/Liber
Box 1205
S-221 05 Lund, Sweden

THE TAPE

The computer tape is distributed by the International Computer Archive of Modern English (ICAME) and can be obtained at cost (see p 34) from

The Norwegian Computing Centre for the Humanities
University of Bergen
P.O. Box 53
N-5014 Bergen, Norway

ORDER FORM: LONDON-LUND CORPUS

To obtain this material you must *enclose a cheque (bank draft, cashier's cheque) with your order made out in Norwegian currency (or the equivalent in English pounds or U.S. dollars)* to: The Norwegian Computing Centre for the Humanities, Bergen, Norway.

Indicate in the table below what you wish to order (prices are given in Norwegian kroner):

Quantity	Magnetic tapes 9 track, 1600 fpi	no. of tapes	price
	London-Lund: text	1 1200 ft.	225
	London-Lund: KWIC I (text categories 1-3, i.e. covering the texts printed in J. Svartvik and R. Quirk (eds.), *A Corpus of English Conversation*, 1980)	1 2400 ft.	350
	London-Lund: KWIC II (text categories 4-12)	1 2400 ft.	350
	Total enclosed		

Name/Institution:

Address:

We understand that the material listed above is the property of the Survey of English Usage and is made available to specialist scholars for the purposes solely of scientific linguistic research, and we agree not to distribute the material or reproduce any part of it for any other purpose than scholarly research. (To be signed by a responsible official of the institution making the order.)

Date:

Signed name:

Print name:

Official position:

Return the signed form to: The Norwegian Computing Centre for the Humanities, P.O. Box 53, N-5014 Bergen-University, Norway.

Description of tape and files

Code:	ASCII
Tracks:	9
Density:	1600 bpi
Label:	None
Parity:	Odd
Files:	1 (information about texts and speakers) + 34 (text categories 1-3 = subgroup A) + 53 (text categories 4-12 = subgroups B-H)
EOF marks:	1 between files, 2 after last file
Record size:	80 bytes
Blocking factor:	100

Description of records

COLS	FORMAT	CONTENTS
1-2	I2	Text category
3-4	I2	Text within category
5	A1	If not blank, subtext within text, indicated by lower case letter
6-8	I3	Slip number (for identification in original London material)
9-12	I4	TU (= tone unit number), essentially increasing (see below)
13	I1	Always 0
14-15	I2	Number of links left in current TU, including current link
16-17	I2	Number of records left in current link, including current record
18	A1	' ' = new speaker, '(' = old speaker continues
19-23	A5	Speaker identity
24	I1	1 = complete link, 2 = incomplete link

COLS	FORMAT	CONTENTS
25	I1	Number of onsets within the TU, not including those occurring in subordinate TUs within this TU
26-27	I2	If not blank, subtext number within text
28-78	A51	Text
79	A1	'/' (right hand margin)
80	A1	Blank

- - - - -

Records make up complete or incomplete TUs. A complete TU is terminated by '#' (number sign). Following '#' there may also be pauses, comments, or semi-colons. Semi-colon (decimal code 59) is used to denote a blank.

Speakers often interrupt each other and, consequently, TUs may be split up into so-called links. TU links may be interwoven. For a TU to be considered properly linked, all links except possibly the last one must be incomplete, ie it lacks '# ', and the speaker should be indicated by '(' in col 18 in all links except possibly the first one. Study TUs 16-17 in text 1-1.

Every TU consists of at least one link. If there are more, links are numbered in decreasing order (cols 14-15).

A link normally goes into one record ('line') but may extend over several. Within each link, therefore, lines have been numbered, in decreasing order.

The printed version of the corpus (see example on p 24), comprising the surreptitiously recorded conversations, has been produced by phototype-setting from the magnetic tape.

3.2 CONCORDANCE

The concordance is divided into two parts: KWIC I, comprising text categories 1-3 = subgroup A, ie covering the texts printed in Svartvik & Quirk 1980, and KWIC II, comprising the remaining texts, ie categories 4-12 = subgroups B-H.

The concordance contains one record for each orthographically transcribed running word in the texts. (See example on p 71.) In addition, phonetically transcribed 'words', enclosed within square brackets [,] as well as contextual comments, enclosed within single parentheses (,), and (sequences of) indistinct words, enclosed within double parentheses ((,)) , have been included, totalling 176 343 instances in KWIC I and 280 749 in KWIC II.

Words uttered by 'non-surreptitious' speakers have not been regarded as keywords in KWIC I, but are included in all contexts surrounding the keywords. This is also true of indications of pause (· or -) in both KWIC I and KWIC II.

The context width is 99 characters.

The sorting order is first by instance type (orthographic, phonetic, comment, indistinct), secondly initial alphabetical, thirdly by prosody, and finally by text order.

Record description

Record length: 168 ASCII characters. Each record consists of a sorting key, a reference to the location in the text, a speaker symbol, and the key word in context.

Sorting key

NAME	CONTENTS	COLUMN	LENGTH
CAPS	Normalized (capitalized) word	1	24
ORTHO	Orthographic word	25	24
DIVC	Word splitting code	49	1
NUCL	Nucleus code	50	1
PROS	Codes for onset, booster, and stress	51	2
PHON	Code for phonetic instance	53	1
COMM	Code for contextual comment	54	1
INDIST	Code for indistinct words	55	1

Location in text

NAME	CONTENTS	COLUMN	LENGTH
CAT	Text category	56	2
TEXT	Text number within category	58	2
SLIP	Slip number in original London material	60	4
TU	Tone unit number within text	64	4
SPI	Speaker identity	68	2

Context

NAME	CONTENTS	COLUMN	LENGTH
KWIC	Key Word In Context	70	99

Fields in the sorting key

CAPS: First alphabetical sorting key, where all letters have been capitalized. Hyphens have been deleted from orthographic words. Phonetic parts of otherwise orthographically transcribed words have also been excluded from this sorting key as well as symbols of prosody and simultaneous talk in all types of units. Pause indications have been deleted except in contextual comments. 'Words' with more than 24 characters have been truncated to that length.

ORTHO: Second alphabetical sorting key. Orthographic words have been normalized with respect to prosody, pause, and simultaneity. Truncated, if needed, to the length of 24 characters.

DIVC: Binary code which takes on the value 1 when an orthographic word is split up into two parts by a pause (eg -sixty- · -five), otherwise 0.

NUCL: (non-zero only in orthographic words)
0 no nucleus
1 fall
2 rise-fall
3 rise
4 fall-rise
5 level

PROS: (non-zero only in orthographic words: contains codes for the 2 first indications of onset, stress, and booster in the word)

0 no indication
1 onset
2 normal stress
3 heavy stress
4 continuance
5 higher than preceding syllable

6 higher than preceding pitch-prominent syllable
7 very high

PHON: Binary code for phonetic type records

COMM: Binary code for contextual comments

INDIST: Binary code for sequences of wholly or partially indistincts words, treated as units

Text location reference fields

CAT: 2-digit text category number with leading space

TEXT: 2-digit text number with leading space fill

SLIP: 4-digit SEU slip number with leading space fill

TU: 4-digit tone unit number with leading space fill

SPI: Speaker identity. Single letter, possibly preceded by (, indicating that the speaker continues where he left off

Context field

KWIC: Key word, surrounded by extra spaces, in context including speaker symbols enclosed within angle brackets <,>. (Word partials, making up words split by pauses, are preceded by hyphens in the context.)

Sorting sequence

The records have been sorted into ascending order on the following fields, from major to minor:

FIELD NAME	SORTING	EXAMPLES
INDIST	Numerical	'a' precedes '((a))' '[a]' precedes '((a))' '(laughs)' precedes '((laughs))'
COMM	Numerical	'laughs' precedes '(laughs)'

PHON	Numerical	'a' precedes '[a]' '((a))' precedes '([[a]])'
CAPS	Alphabetical	'a' precedes 'BBC'
ORTHO	Alphabetical	'Stone' precedes 'stone' 'it's' precedes 'its' 'flower-pot' precedes 'flowerpot'
DIVC	Numerical	'remember' precedes '-re · -member'
NUCL	Numerical	'too' precedes 't\oo' 't\/oo' precedes 't=oo'
PROS	Numerical	'too' precedes ':too' '"_t\oo' precedes '"|t\oo'
		Other things being equal:
CAT	Numerical	Category 1 precedes category 2
TEXT	Numerical	Text 1.1 precedes text 1.2
TU	Numerical	Tone unit 1 precedes tone unit 2

Collating sequence

The following collating sequence has been used in the alphabetical sorting: 1) space 2) !"# $%&' *+,-./ (ASCII 33-39, 42-47) 3) 0-9 (ASCII 48-57) 4) :;<=>?@^ (ASCII 58-64, 96) 5) AaBb...Zz (ASCII 65,97,66,98,...,90,122) 6) [\]^ _ {|}~ ■ (ASCII 91-95, 123-127) 7) () (ASCII 40-41).

3.3 WORD-LISTS

Edited frequency word lists based on the concordance are being prepared. They will be sorted both alphabetically and by descending frequency. See p 43f for a list of the 100 most frequent words in the London-Lund corpus of spoken English

arranged in Table 4 by descending frequency. ('Rk' = rank; 'distr' = distribution, ie number of texts in which the respective words occur; 'n' = total number of texts). The words are listed according to rank order for the whole corpus (subcategories A-H) and also rank order and frequencies for subcategory A and subcategories B-H separately.

Table 5 gives the 100 most frequent words listed according to rank order in subcategory A and, for comparison, the respective rank orders for subcategories B-H and also LOB (the Lancaster-Oslo/Bergen Corpus) and Brown (the Brown Corpus). '-' denotes that the word is not among the 100 most frequent words in the respective corpus.

Subsequent more comprehensive and refined analyses will provide better information about the vocabulary of spoken English, but even the lists in Tables 4 and 5 give some interesting insights (compare, for example, the rank order of you, yes, well, know, think and sort in Table 5).

TABLE 4: THE 100 MOST FREQUENT WORDS IN THE LONDON-LUND CORPUS

WORD	WHOLE CORPUS (A-H)	SUBGROUP A			SUBGROUPS B-H		
	rk	rk	freq	distr n =34	rk	freq	distr n=53
the	1	2	5,833	34	1	11,904	53
and	2	3	4,825	34	2	7,874	53
I	3	1	5,920	34	5	6,489	53
to	4	4	3,830	34	3	7,016	53
of	5	6	3,647	34	4	6,624	53
a	6	7	3,453	34	6	5,550	53
you	7	5	3,776	34	9	4,233	51
that	8	10	2,613	34	7	4,578	53
in	9	9	2,650	34	8	4,523	53
it	10	8	3,094	34	10	3,900	53
is	11	13	1,773	34	11	2,980	53
yes	12	11	2,099	34	14	2,164	43
was	13	12	1,789	34	12	2,327	53
this	14	17	1,350	34	13	2,189	53
but	15	16	1,401	34	16	1,997	53
on	16	20	1,187	34	15	2,106	53
well	17	14	1,432	34	22	1,671	52
he	18	18	1,315	34	21	1,687	53
have	19	23	1,076	34	20	1,732	53
for	20	27	972	34	19	1,780	53
be	21	28	936	34	17	1,799	53
know	22	15	1,429	34	28	1,287	51
we	23	29	894	34	18	1,790	53
they	24	19	1,198	34	26	1,462	53
it's	25	21	1,162	34	27	1,401	52
at	26	33	873	34	24	1,496	53
[ə:m]	27	22	1,121	34	33	1,207	42
think	28	25	1,042	34	30	1,234	53
as	29	40	765	34	23	1,506	53

Table 4 continued

WORD	WHOLE CORPUS (A-H)	SUBGROUP A			SUBGROUPS B-H		
	rk	rk	freq	distr n=34	rk	freq	distr n=53
so	30	24	1,050	34	32	1,220	53
no	31	25	1,042	34	37	1,153	53
with	32	42	722	34	25	1,469	53
all	33	29	894	34	33	1,207	53
very	34	35	836	34	31	1,232	53
not	35	38	778	34	29	1,248	53
one	36	36	831	34	36	1,179	53
what	37	32	875	34	41	1,098	52
there	38	41	748	34	35	1,205	53
about	39	37	791	34	42	1,005	53
do	40	34	870	34	47	917	53
if	41	48	651	34	38	1,113	53
are	42	49	639	34	40	1,104	53
which	43	55	563	34	39	1,107	53
or	44	47	662	34	43	996	52
[m]	45	31	880	33	58	741	28
yeah	46	50	610	31	46	929	31
oh	47	43	716	34	52	803	42
don't	48	39	767	34	61	737	51
got	49	45	674	34	50	816	52
see	50	45	674	34	60	739	53
now	51	68	452	34	44	958	53
just	52	54	577	34	49	832	53
from	53	69	450	34	45	936	53
would	54	51	595	34	56	772	52
had	55	57	542	34	50	816	52
because	56	58	541	34	63	708	53
I'm	57	56	548	34	65	694	48
my	58	66	466	34	55	773	51
she	59	72	436	27	53	802	40
mean	60	44	713	33	89	510	44
up	61	63	487	34	62	729	53

WORD	WHOLE CORPUS (A-H)	SUBGROUP A			SUBGROUPS B-H		
	rk	rk	freq	distr n=34	rk	freq	distr n=53
that's	62	53	583	34	72	631	48
then	63	61	517	34	67	692	53
were	64	69	450	34	58	741	52
by	65	99	286	34	48	898	53
me	66	67	453	34	68	685	50
when	67	64	469	34	71	666	53
been	68	78	406	34	64	705	53
out	69	75	422	34	69	684	53
can	70	76	408	34	66	693	53
like	71	62	500	34	76	589	52
going	72	80	404	34	70	675	53
right	73	97	287	34	54	783	52
get	74	64	469	34	76	589	52
[əm]	75	74	432	31	73	625	42
said	76	60	534	34	87	518	47
them	77	73	433	34	75	618	53
really	78	59	539	34	90	503	52
sort	79	52	590	34	101	432	43
people	80	71	444	34	82	538	51
an	81	76	408	34	84	531	52
two	81	91	320	34	74	619	52
who	83	93	315	33	79	574	53
time	84	87	353	34	86	519	53
his	85	94	311	33	57	753	50
go	86	83	364	34	94	484	53
say	87	88	349	33	92	495	52
thing	88	78	406	34	99	436	48
him	89	91	320	34	88	512	52
some	90	86	360	34	95	467	53
quite	91	81	391	34	102	430	49
much	91	89	329	34	93	492	53
your	93	103	275	33	80	541	44

Table 4 continued

WORD	WHOLE CORPUS (A-H)	SUBGROUP A			SUBGROUPS B-H		
	rk	rk	freq	distr n=34	rk	freq	distr n=53
has	94	106	260	34	82	538	53
I've	95	82	369	34	105	417	45
her	96	113	235	26	85	523	38
more	97	97	287	34	97	461	53
did	98	90	322	34	104	424	50
[ði:]	99	122	204	31	80	541	52
he's	100	84	363	34	115	377	47

TABLE 5: THE 100 MOST FREQUENT WORDS IN THE LONDON-LUND CORPUS COMPARED WITH LOB AND BROWN

(Frequency based on subgroup A)

WORD	LONDON-LUND CORPUS		LOB	BROWN
	SUBGROUP A	SUBGROUPS B-H		
I	1	5	17	20
the	2	1	1	1
and	3	2	3	3
to	4	3	4	4
you	5	9	32	33
of	6	4	2	2
a	7	6	5	5
it	8	10	10	12
in	9	8	6	6
that	10	7	7	7
yes	11	14	-	-
was	12	12	9	9
is	13	11	8	8
well	14	22	95	100
know	15	28	-	-
but	16	16	24	25
this	17	13	22	21
he	18	21	12	10
they	19	26	33	30
on	20	15	16	16
it's	21	27	-	-
[ə:m]	22	33	-	-
have	23	20	26	28
so	24	32	46	52
no	25	37	47	49
think	25	30	-	-
for	27	19	11	11
be	28	17	15	17
we	29	18	40	41
all	29	33	39	36
[m̩]	31	58	-	-

Table 5 continued

WORD	LONDON-LUND CORPUS		LOB	BROWN
	SUBGROUP A	SUBGROUPS B-H		
what	32	41	55	54
at	33	24	19	18
do	34	47	73	72
very	35	31	81	-
one	36	36	38	32
about	37	42	54	57
not	38	29	23	23
don't	39	61	-	-
as	40	23	13	14
there	41	35	36	38
with	42	25	14	13
oh	43	52	-	-
mean	44	89	-	-
got	45	50	-	-
see	45	60	-	-
or	47	43	31	27
if	48	38	45	50
are	49	40	27	24
yeah	50	46	-	-
would	51	56	43	39
sort	52	-	-	-
that's	53	72	-	-
just	54	49	-	-
which	55	39	28	31
I'm	56	65	-	-
had	57	50	21	22
because	58	63	-	-
really	59	90	-	-
said	60	87	52	53
then	61	67	68	71
like	62	76	83	78
up	63	62	56	55
when	64	71	44	45

Table 5 continued

WORD	LONDON-LUND CORPUS		LOB	BROWN
	SUBGROUP A	SUBGROUPS B-H		
get	64	76	-	-
my	66	55	58	75
me	67	68	66	82
now	68	44	72	76
from	69	45	25	26
were	69	58	35	34
people	71	82	100	-
she	72	53	30	37
them	73	75	60	59
[əm]	74	73	-	-
out	75	69	53	51
can	76	66	61	61
an	76	84	34	29
thing	78	99	-	-
been	78	64	37	43
going	80	70	-	-
quite	81	-	-	-
I've	82	-	-	-
go	83	94	-	-
he's	84	-	-	-
[m̀]	84	-	-	-
some	86	95	57	65
time	87	86	63	66
say	88	92	-	-
much	89	93	89	97
did	90	-	-	88
him	91	88	49	42
two	91	74	67	69
who	93	79	50	46
his	94	57	18	15
good	95	-	-	-
something	96	-	-	-
right	97	54	-	-
more	97	97	51	48
by	99	48	20	19
any	99	-	75	74

4 Research Based on the London-Lund Corpus

The London-Lund Corpus has been and is being used for various studies on, for example, grammatical tagging, reference, questions and responses, negation, turn-taking and interruption.

4.1 WORD-CLASS TAGGING

By 'tagging' we mean the assignment of a lexical-grammatical description to linguistic units in the texts. We will here distinguish between word-class tagging and syntactic tagging.

Principles of coding

The basic aim of the word-class tagging of the London-Lund Corpus agrees with that of the Brown Corpus: 'supplying each word in the Corpus with a symbol indicating its place in a taxonomy based on surface syntactic function' (Francis 1980:198). Although our choice of word-class labels differs to some extent from the labels used for the Brown and LOB corpora, the systems are probably compatible enough to allow comparison of all three corpora.

Among the principles for our word-class tagging are the following:

- The aim is not to tag the whole corpus but only parts of it that are of special interest to our research, eg a subset of texts representing different speech situations such as informal conversation, public debates, prepared oration, commentary and news broadcast.
- The word-class tagging is not considered an end in itself but rather a stepping-stone to syntactic analysis. For this reason we want to have a system that is as open-ended as possible for renewal of connection in the light of our results in the syntactic analysis. We do not expect the currently adopted subcategorization of

word-classes to be the final one; rather it should be flexible reflecting both specific personal research interests and general usefulness for the higher level analysis.

- The tagging is designed as an interactive, semi-automatic process, partly to save time, partly to enable us to improve the rules in the course of the tagging.
- The tags are quite traditional and based on a mixture of formal and functional criteria. With computer techniques it is easy to relabel tags for categories and words.

Word-class tagging involves the assignment of word-class markers (tags) to all graphic words in a text, eg in tone units 546-643 of Text S.1.1, which is an informal conversation between two university teachers. Fig 4 shows the printed version using a photosetting technique. Fig 5 shows a hard copy output of the same text portion using a Facit printer with a large character set including certain prosodic and phonetic symbols; grammatical tags are given in angle brackets. The Appendix lists the word-class tags used in Fig 5. Here is an example from this text (tone units 553-554):

LEXICAL WORD	TAG	LEXICAL-GRAMMATICAL CATEGORY
this	RG	pronoun/demonstrative/singular
is	VB+3	verb/*be*/3rd person/sg/present
to	PD	infinitive marker *to*
run	VA+Ø	verb/main/base form
their	TB	determiner/possessive
coordinating	VA+G	verb/main/*-ing* form
machinery	NC	noun/common/singular
you	RC	pronoun/personal/unmarked for case
know	VA+Ø	verb/main/base form

Word-class tags are valuable for several reasons:

- For homograph separation, eg to distinguish *to* as infinitive marker (PD) eg *to run* from *to* as preposition (PA), eg *to Burgos*;
- For lemmatization, eg bringing together different forms of the lexeme *be*, indicated by VB followed by a '+' sign and a code for the different forms of the verb: *be* = VB+Ø, *is* = VB+3, *been* = VB+N, *being* = VB+G;
- For syntactic analysis, eg in marking the string of tagged words *their* (TB) *coordinating* (VA+G) *machinery* (NC) as having the structure of a noun phrase (see the phrase analysis on p 76).

Our word-class tags consist of two capital letters, of which the first denotes a major category, such as 'N' for 'noun' and 'V' for 'verb', and the latter a subcategory. Thus: 'VA' for 'main verb', 'VM' for 'modal verb'. The reason for this hierarchical order is that it simplifies the writing of syntactic parsing rules. In addition to these two letters we sometimes use '+' for special forms, eg '+2' for plural, '+Z' for genitive, '+Ø' for the base form of a verb, '+D' for the past form and '+G' for the *-ing* form; '*' denotes contracted forms, thus 'VB+3' for *is*, 'AN' for *not* and 'VB+3*AN' for *isn't*.

First design of an interactive tagging program

Our original plan was to use a general-purpose dictionary, *Longman Dictionary of Contemporary English (LDOCE)* as our basic source of tagging information. Word-class labels taken from *LDOCE* and stored on magnetic tape would be assigned to every word type in our texts, translated into our own system of codes, and transferred to each word token in the texts. Disambiguation context programs would be applied to cases of homography (multiple tags). Finally, all the word tags would be checked and corrected manually on the terminal screen, enabling an assessment of the scope and limitations of the fully automatic tagging phase.

FIGURE 4. PRINTED VERSION OF TEXT S.1.1
(TONE UNITS 546-643)

B 546 I'm ‖going to BŪRGOS‖ · 547 ‖WĒDNESDAY‖ · 548 ‖WĒEK‖ · 549 ‖YĒS‖
550 the ‖TĒNTH‖ · 551 ‖coming back on the TWENTY-NĪNTH‖

A 552 and ‖this is to give △VĪVAS {‖THÉRE‖} ‖ -

B 553 [ɔ:] ‖this is to · ▵run their coordinating MACHĪNE«RY»‖ - 554 ‖you
KNŌW‖ 555 ‖to · establish the ▵STĀNDARDS‖ 556 and ‖all THĀT ▷kind of
▷thing‖ - - -

A 557 ‖have you △MĒT our man Yoolet YĒT‖ - 558 ☆«the ‖one who's a △student
for the · »☆ 'DIPLŌMA‖

B 559 ☆[ɔ:] ‖NŌ‖ 560 «‖NŌ‖»☆ 561 ‖NŌ‖

A 562 ‖[mhḿ]‖

B 563 I ‖knew that he was CŌMING‖ 564 I've ‖heard Stan △Carter MĒNTION
▷him‖ 565 and ☆ · «they've ‖obviously»☆

A 566 ☆«‖YĒAH‖» · 567 ‖now that☆ he's finished his EXAMIN▵ĀTIONS‖ 568 you
‖KNŌW‖ 569 ‖you'll be SĒEING him‖ 570 at ‖[ðəi] at ‖[ðəi] · V̂IVA‖ · 571 ‖on
WĒDNESDAY‖

B 572 [ə] ‖YĒS‖ 573 ‖[hm hm̄]‖

A 574 [ɔ:m] I'm ‖sure that - «you know» he would be ▵awfully ▵GRĀTEFUL‖
575 if you could - ‖see him in your △ŌFFICE SŌMETIME‖ ·

B 576 ‖well I'd ▵like to have a chat with [ú▷pɔ▷te:]‖

A 577 «2 sylls» because ‖if · ‖if he · DŌESN'T ▷work‖ 578 in ‖close ·
collabor▵ation with YŌU‖ 579 and - - and ‖try to get - your EX△PĒRIENCE‖ -
- 580 he's going to go ‖badly · at △SĒA‖ · 581 cos ‖he's going to have a lot
of RESPONSI▵BĪLITY «on HIMSĒLF‖ 582 ☆1 to 2 sylls»☆

B 583 ☆I ‖think they're☆ ▵grooming him for - - ▵quite a a RE▵SPŌNSIBLE ☆JŌB‖
584 at · ☆ MAIDRĪD‖ - - 585 I

A 586 ☆‖[m̄]‖ · 587 ‖[m̄]☆

>B 585 ‖heard his name ▵mentioned by - ▵CĀRTER‖ 588 «I ‖THĪNK‖» 589 ‖by
DĀRLINGTON‖ 590 ‖while I was △DŌWN ▷there‖ - - -

A 591 ‖DĪD ▷you‖ - - -

B 592 I ‖think they've got quite a good O△PĪNION of him‖ -

A 593 ‖well [ə] ▵I ΔI have TÒO‖

B 594 ‖[ḿ]‖

A 595 [ɔ:m] - - «‖well I mean» · the ‖way these chaps ΔGÒ‖

B 596 [m]‖[hm̀]‖ - -

A 597 [ə:] he's he's a ‖THÒUGHTFUL MÁN‖ 598 he's ‖not a ΔHǍSTY ▷man‖ - - -
599 ☆«‖I]»☆

B 600 ☆«‖[m̀]‖»☆ - - 601 «‖I've got» a 'PRÒBLEM for you my LÁD‖ - -

A 602 a ‖PRÒBLEM‖

B 603 ‖YÈS‖

A 604 «‖I'm at your ▵SÈRVICE‖» - - - 605 this is an‖other of your
'ΔBǍWDINESSES I'll BÉT‖ -

B 606 [əhɔ:] - - it's ‖not worth · the TRÒUBLE‖ - - 607 ‖how do you ǍNALYSE‖
- 608 ‖WÔRTH‖ · 609 it's ‖not 'worth the TRÒUBLE‖ - - - 610 ‖not 'worth the
TRÒUBLE‖ - - - 611 ‖this is [ə:] «the» ▵little · ▵fellow «[æ] in» the
department of sty▵listics at · ▵BÙRGOS‖ · 612 «‖[bɔ̀ʌdi]‖» · 613 and he
‖wasn't ▵quite · ▷certain how to ▵TǍCKLE‖ · 614 ‖worth the ΔTRÒUBLE‖
615 in ‖that sort of «COLLO▵CǍTION»‖ - - -

A 616 ‖good QUÈSTION‖ - - 617 and it's ‖NÒT one of your {‖BǍ̂WDINESSES‖} ‖ ·
618 I ‖take it all ▵BǍCK‖

B 619 (- laughs)

A 620 «3 ‖sylls to YÒU‖» - 621 [ɔ:m] - - - ‖oh it's Δ{SÙI} GÈNERIS‖ -
622 ‖ÎSN'T it‖ - - 623 ☆«it»☆ ‖doesn't pattern with ▵ǍNYTHING‖

B 624 ☆‖YÈS‖☆ 625 well ‖that's the whole ▵TRÒUBLE « {you ‖SÊE‖} ‖» 626 that's
‖obviously what is worrying ΔHÌM‖ - - -

A 627 [ɔ:m] - - - it ‖seems to MÈ‖ 628 that - [ɔ:m] - - -

B 629 there ‖are similar ▵PǍTTERNS‖ 630 «[ə] I mean» ‖one would say for
example it's ▵not worth ΔRÈADING‖ · 631 ☆«which is similar to»☆ it is ‖not

A 632 ☆«‖ǍH‖ 633 ‖YÈS‖ 634 ‖YÈS‖ 635 ‖YÈS‖»☆

>B 631 worth · [ΔðÎ:] · {RÈADING‖} ‖

A 636 ‖YÈS‖ · 637 «well» ‖not worth ΔNÌNEPENCE‖ 638 ‖YÊS‖ 639 ‖what I
▵mean ÎS‖ 640 «[ð]» the ‖whole point here is WÔRTH‖ · 641 ‖ÎSN'T it‖

B 642 ‖that's «the [w]» ‖that's the word he's ☆ǍFTER‖ 643 ‖YÈS‖☆

FIGURE 5. PRINTOUT OF A TAGLED TEXT (TEXT S.1.1, TONE UNITS 546-643)


```
0101000546 B      I'm<RA▲VB+1> ‖going<VA+G> to<PA> B→urgos■<NP> ·
0101000547 B      ‖W→ednesday■<NP> ·
0101000548 B      ‖w↘eek■<NC> ·
0101000549 B      ‖y↘es■<AS>
0101000550 B      the<TA> ‖t↘enth■<JQ> ·
0101000551 B      ‖coming<VA+G> back<AC> on<PA> the<TA> twenty-n↘inth■<JQ>
0101000552 A      and<CA> ‖this<RG> is<VB+3> to<PD> give<VA+O> !v↗ivas<NC+2>
                  {‖th↗ere■}■<AC> -
0101000553 B      [ə:]<XR> ‖this<RG> is<VB+3> to<PD> · :run<VA+O>
                  their<TB> coordinating<VA+G> mach↘ine«ry»■ -
0101000554 B      ‖you<RC> kn↗ow■<VA+O>
0101000555 B      ‖to<PD> · establish<VA+O> the<TA> :st↘andards■<NC+2>
0101000556 B      and<CA> ‖all<EC> th↘at<TD> ▷kind<NC> of<PA> ▷thing■<NC>
                  - - -
0101000557 A      ‖have<VH+O> you<RC> !m↘et<VA+N> our<TB> man<NC>
                  Yoolet<NP> y↗et■<AC> -
0101000558 A      ▲«the ‖one who's a !student for the ·
                  »▲.dipl↘oma■<NC>
0101000559 B      ▲[ə:]<XR> ‖n↘o■<AR>
0101000560 B      «‖n↘o■»▲
0101000561 B      ‖n↘o■<AR>
0101000562 A      ‖[mh↗m]■<XM>
```

```
0101000563 B      I<RA> ‖knew<VA+0> that<CD> he<RA> was<VB+5>
                  c^oming■<VA+G>
0101000564 B      I've<RA*VH+0> ‖heard<VA+N> Stan<NP> !Carter<NP>
                  m^ention<VA+0> ▷him■<RB>
0101000565 B      and<CA> *· «they've ‖obviously»*
0101000566 A      *«‖y^eah■» ·
0101000567 A      ‖now<AC> that*<GD> he's<RA*VH+3> finished<VA+N> his<TB>
                  examin:^ations■<NC+2>
0101000568 A      you<RC> ‖kn^ow■<VA+0>
0101000569 A      ‖you'll<RC*VM+8> be<VB+0> s^eeing<VA+G> him■<RB>
0101000570 A      at<PA> ‖[δi]<TA> at<PA> ‖[δi]<TA> · v^iva■<NC> ·
0101000571 A      ‖on<PA> W^ednesday■<NP>
0101000572 B      [ə]<XR> ‖y^es■<AS>
0101000573 B      ‖[hm h+m]■
0101000574 A      [ə:m]<XR> I'm<RA*VB+1> ‖sure<JA> that<CD> - «you
                  know» he<RA> would<VM+9> be<VB+0> :awfully<AW>
                  :gr^ateful■<JA>
0101000575 A      if<CC> you<RC> could<VM+9> - ‖see<VA+0> him<RB> in<PA>
                  your<TB> !^office<NC> s^ometime■<AC> ·
0101000576 B      ‖well<IA> I'd<RA*VM+9> :like<VA+0> to<PD> have<VH+0>
                  a<TF> chat<NC> with<PA> [^u▷po▷te:]■
0101000577 A      «2 sylls» because<CC> ‖if<CC> · ‖if<CC> he<RA> ·
```

57

Figure 5 continued


```
                d^oesn't<VD+3*AN> ▷work■<VA+0>
0101000578 A    in<PA> ‖close<JA> · collabor:ation<NC> with<PA>
                y→ou■<RC>
0101000579 A    and<CA> - - and<CA> ‖try<VA+0> to<PD> get<VA+0> -
                your<TB> ex!p↘erience■<NC> - -
0101000580 A    he's<RA*VB+3> going<VA+G> to<PD> go<VA+0> ‖badly<AW>
                · at<PA> !s↘ea■<NC> ·
0101000581 A    cos<CC> ‖he's<RA*VB+3> going<VA+G> to<PD> have<VH+0>
                a<TF> lot<NC> of<PA> responsi:b↘ility<NC> «on
                hims↗elf■
0101000582 A    *1 to 2 sylls»*
0101000583 B    *I<RA> ‖think<VA+0> they're*<RA*VB+4> :grooming<VA+G>
                him<RB> for<PA> - - :quite<EL> a<TF> a<TF>
                re:sp↘onsible<JA> *j↗ob■<NC>
0101000584 B    at<PA> ·* Ma‖dr↗id■<NP> - -
0101000585 B    I<RA> ‖heard<VA+0> his<TB> name<NC> :mentioned<VA+N>
                by<PA> - :C↘arter■<NP>
0101000586 A    *‖[→m]■<XM> ·
0101000587 A    ‖[→m]*<XM>
0101000588 B    «I ‖th↗ink■»
0101000589 B    ‖by<PA> D↘arlington■<NP>
0101000590 B    ‖while<CC> I<RA> was<VB+5> !d↘own<AC> ▷there■<AC> -
                - -
```

```
0101000591 A      ‖d↘id<VD+0> ▷you■<RC> - - -
0101000592 B      I<RA> ‖think<VA+0> they've<RA★VH+0> got<VA+N> quite<EL>
                  a<TF> good<JA> o!p↘inion<NC> of<PA> him■<RB> -
0101000593 A      ‖well<IA> [ə]<XR> :I<RA> !I<RA> have<VH+0> t↘oo■<AT>
0101000594 B      ‖[↗m]■<XM>
0101000595 A      [ə:m]<XR> - - «‖well I mean» · the<TA> ‖way<NC>
                  these<TE> chaps<NC+2> !g↘o■<VA+0>
0101000596 B      [m]‖[h↘m]■<XM> - -
0101000597 A      [ə:]<XR> he↷s<RA★VB+3> he's<RA★VB+3> a<TF>
                  ‖th↘oughtful<JA> m↗an■<NC>
0101000598 A      he's<RA★VB+3> ‖not<AN> a<TF> !h↷asty<JA> ▷man■<NC>
                  - - -
0101000599 A      ★«‖I»★
0101000600 B      ★«‖[↘m]■»★ - -
0101000601 B      «‖I've got» a<TF> .pr↘oblem<NC> for<PA> you<RC> my<TB>
                  l↗ad■<NC> - -
0101000602 A      a<TF> ‖pr↷oblem■<NC>
0101000603 B      ‖y↘es■<AS>
0101000604 A      «‖I'm at your :s↘ervice■» - - -
0101000605 A      this<RG> is<VB+3> an‖other<RP> of<PA> your<TB>
                  .!b↘awdinesses<NC+2> I'll<RA★VM+8> b↗et■<VA+0> -
0101000606 B      [əhə:] - - it's<RC★VB+3> ‖not<AN> worth<PA> · the<TA>
```

Figure 5 continued

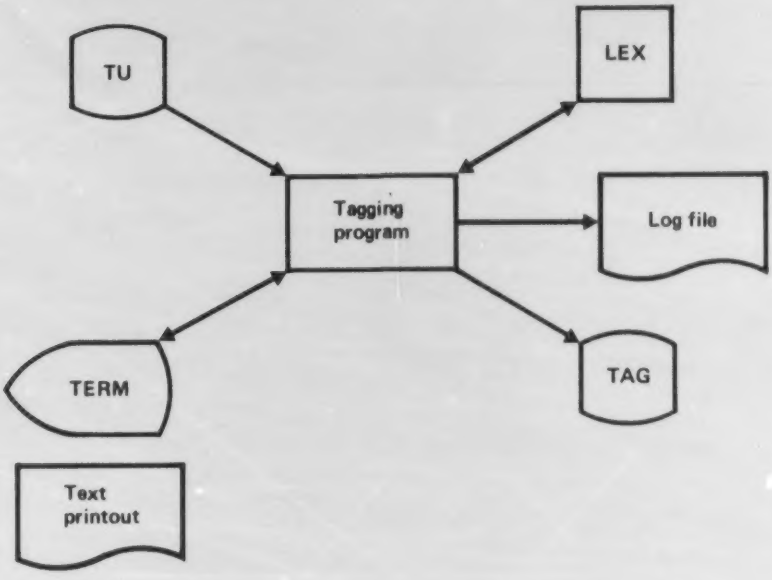
```
                tr↘ouble■<NC> - -
0101000607 B    ‖how<AC> do<VD+0> you<RC> ↘analyse■<VA+0> -
0101000608 B    ‖w↗orth■<PA> .
0101000609 B    it's<RC★VB+3> ‖not<AN> .worth<PA> the<TA> tr↘ouble■<NC>
                - - -
0101000610 B    ‖not<AN> .worth<PA> the<TA> tr↘ouble■<NC> - - -
0101000611 B    ‖this<RG> is<VB+3> [ə:]<XR> «the» :little<JA> .
                :fellow<NC> «[ae] in» the<TA> department<NC> of<PA>
                sty:listics<NC> at<PA> . :B↘urgos■<NP> .
0101000612 B    «‖[b↘oadi]■» .
0101000613 B    and<CA> he<RA> ‖wasn't<VB+5★AN> :quite<AC> .
                ▷certain<JA> how<AB> to<PD> :t↘ackle■<VA+0> .
0101000614 B    ‖worth<PA> the<TA> !tr↘ouble■<NC>
0101000615 B    in<PA> ‖that<TD> sort<NC> of<PA> «collo:c↘ation»■
                - - -
0101000616 A    ‖good<JA> qu↘estion■<NC> - -
0101000617 A    and<CA> it's<RC★VB+3> ‖n↘ot<AN> one<RJ> of<PA> your<TB>
                (‖b→awdinesses■)■<NC+2> .
0101000618 A    I<RA> ‖take<VA+0> it<RC> all<RP> :b↘ack■<AC>
0101000619 B    ( - laughs)
0101000620 A    «3 ‖sylls to y↘ou■» -
0101000621 A    [ə:m]<XR> - - - ‖oh<IA> it's<RC★VB+3> !(s↘ui)<XX>
                g↘eneris■<XX> -
```

```
0101000622 A     ‖↘isn't<VB+3★AN> it■<RC> - -
0101000623 A     ★«it»★ ‖doesn't<VD+3★AN> pattern<VA+0> with<PA>
                 :↘anything■<RM>
0101000624 B     ★‖y↘es■★<AS>
0101000625 B     well<IA> ‖that's<RD★VB+3> the<TA> whole<JA>
                 :tr↘ouble<NC> «{you ‖s→ee■}■»
0101000626 B     that's<RD★VB+3> ‖obviously<AW> what<GC> is<VB+3>
                 worrying<VA+G> !h↘im■<RB> - - -
0101000627 A     [ə:m]<XR> - - - it<RC> ‖seems<VA+3> to<PA> m↘e■<RB>
0101000628 A     that<CD> - [ə:m]<XR> - - -
0101000629 B     there<AX> ‖are<VB+4> similar<JA> :p↘atterns■<NC+2>
0101000630 B     «[ə] I mean» ‖one<RJ> would<VM+9> say<VA+0> for<PA>
                 example<NC> it's<RC★VB+3> :not<AN> worth<PA>
                 !r↘eading■<VA+G> ·
0101000631 B     ★«which is similar to»★ it<RC> is<VB+3> ‖not<AN>
                 worth<PA> · [!ð↘i:]<TA> · {r↘eading■}■<NC>
0101000632 A     ★«‖↘ah■
0101000633 A     ‖y↘es■
0101000634 A     ‖y↘es■
0101000635 A     ‖y↘es■»★
0101000636 A     ‖y↘es■<AS> ·
0101000637 A     «well» ‖not<AN> worth<PA> !n↘inepence■<NC>
```

Figure 5 continued

```
0101000638 A      ‖y→es■<AS>
0101000639 A      ‖what<GC> I<RA> :mean<VA+0> ↗is■<VB+3>
0101000640 A      «[ð]» the<TA> ‖whole<JA> point<NC> here<AC> is<VB+3>
                  w^orth■<PA> .
0101000641 A      ‖^isn't<VB+3★AN> it■<RC>
0101000642 B      ‖that's<RD★VB+3> «the [w]» ‖that's<RD★VB+3> the<TA>
                  word<NC> he's<RA★VB+3> ★^after■<PA>
0101000643 B      ‖y↘es■★<AS>
```

FIGURE 6. INTERACTIVE TAGGING SYSTEM DESIGN



LEX: Dynamic lexicon in core storage
TU: Texts stored for direct access as tone unit records
TAG: Tone unit texts with tags added
TERM: Linguist working at terminal

One attractive feature of this plan for a dictionary-based approach was that it would have made the detailed information contained in the grammatical codes of *LDOCE* available for subsequent phrase tagging. Another attractive feature was that it would have permitted us to make interesting comparisons between the vocabulary of *LDOCE* and that of our texts. Also, an important part of the technical implementation of this plan was the storing of the texts in a linked, highly interconnected network, facilitating systematic searching and updating.

Final design

For various reasons, mainly practical, this plan had to be abandoned in favour of a simpler approach which had been tested on a small scale parallel to the preparations for the dictionary-based alternative. This preparatory work proved extremely time-consuming. When we had got physical access to the *LDOCE* magnetic tape, converted it into a form palatable to our UNIVAC computer, and extracted the relevant information, much work still remained to be done, including frequent manual intervention.

In view of the fact that all these elaborate preparations would still have left us with the need for a final complete manual check, we finally settled for the more direct systems solution sketched in Figure 6. (The work done on adopting *LDOCE* was not, however, entirely a waste of time. It enabled us to make some interesting selective printouts of certain grammatical classes.)

In the final approach the texts were tagged token by token on the terminal screen in a highly interactive way by a linguist aided by a small high-frequency lexicon and some heuristics incorporated into the tagging program, written in the Simula programming language. This program presents the

tone units one by one, first in full, including tone unit number, speaker symbol and prosodic information. At this point the linguist sitting at the terminal can choose between terminating the tagging session, skipping the current tone unit, or going on to tag it, adding in passing any comment he likes to a special log file. If he decides to go on, the tone unit is presented as a sequence of orthographic words, each written on a separate line and followed by a suggested tag, which the linguist can okay by pressing carriage return, or correct by typing in the proper alternative.

Output from the program

The main output of the tagging program is simply a copy of the input text, where each word is immediately followed by its word-class tag enclosed within angle brackets. This simple-minded copying is admittedly a clumsier approach than the integrated network storage envisaged in the first design sketch. It involves duplication of storage, thereby making updating more difficult. Our solution was much cheaper, however, since only a small fraction of the texts, and not an entire data base needed to be directly accessible to the tagging program at any one time.

The program also outputs the log file mentioned above which contains the linguist's comments and observations. Further logging information records the time of the beginning and end of every tagging session and describes which tone units and texts were processed and what errors, if any, were found.

Input to the program

The program accepts input from three sources: from the linguist sitting at the terminal, from the lexicon described in the next section, and, naturally, from the texts to be tagged. Before being presented on the screen, the original

texts had to be somewhat reformatted. Since the tone unit is our basic unit of analysis (see section 4.2), especially in phrase analysis, and the conversations contain numerous tone units split up into 'links', sometimes interwoven in a rather complicated manner when the speakers interrupt each other, the split tone units had to be reassembled. This preparatory work could be done fully automatically. The records in the resulting tone unit file contain, for each tone unit, a tone unit number, references to the category and text/subtext where the tone unit is found, an indication of the speaker and the text proper with prosodic information.

During the terminal work the linguists had access to a detailed tagging manual as well as a complete printout of the text being tagged. The printout was consulted in the comparatively few cases when the tone unit was insufficient as a context for word-class tagging.

The lexicon

The most important heuristic information used by the tagging program is provided by the dynamic high frequency lexicon stored in the computer. The basic tagging strategy is to look up each input word in the lexicon and assign to it the most frequently used tag found under that entry. This simple strategy is quite efficient. From the statistical study of vocabulary it is well known that a small number of frequent word types make up the bulk of any text. For example, the 100 most frequent words account for nearly every second word of the Brown corpus.

The program uses the lexicon in a dynamic way, adding to it all 'new' words encountered in the texts. For each lexical entry it maintains the cumulative total frequency as well as a list of possible tags, each one with its corresponding partial frequency. In this way the statistically most probable tag can be chosen.

Originally the lexicon contained about 400 entries whose tags had been coded manually. It was then allowed to grow freely until it reached the size of about 1000 entries, when it became too large to go into core storage. From then on a small pruning program had to be applied at regular intervals, deleting the least frequent entries (and all proper names). This simple pruning policy, which pays attention to total frequency only, disregarding the dispersion of words over the texts, is probably not optimal. For instance, it was observed that certain highly topic-dependent words clustered in one single text seemed to stay forever in the lexicon, while some seemingly frequent and well dispersed pronouns were repeatedly thrown out, never managing to establish themselves as permanent members of the lexicon.

On the whole, however, the dynamic lexicon performed well, making effective use of the repetitive nature of the vocabulary of the conversations.

Suffix and context frame rules

The lexicon is the source of information that provides by far the largest number of correctly predicted tags. The remaining heuristic rules are few in number and their contribution to the overall performance of the tagging program is small. They have been included mainly to illustrate the possibility of using such information. More rules can, however, easily be added to the program.

These rules include, for example, a special heuristic for the word *that*, a simple context rule for infinitives, and a set of graphic pattern-matching ('suffix stripping') rules which are applied to words that can neither be found in the lexicon, nor handled by the other heuristic rules. If all else fails, the default action is to predict that the unknown word is a singular noun.

The *that* heuristic takes advantage of the prosodic information for the disambiguation of the four functions of this word. A pilot study based on 442 occurrences gave the distribution shown in Table 6.

TABLE 6. DISTRIBUTION OF PROSODY AND GRAMMATICAL FUNCTIONS OF *THAT*

	+prosody	-prosody	Total
Relative pronoun	0	52	52
Conjunction	7	185	192
Determiner	26	16	42
Demonstrative pronoun	139	17	156
Total	172	270	442

'Prosodic information' includes presence/absence of nucleus, stress, onset or booster. Occurrences of *that* with prosodic marking are classified as determiners or demonstrative pronouns, depending on whether or not there is an immediately preceding noun. By the same criterion, prosodically unmarked occurrences are separated into relative pronouns and subordinating conjunctions.

On the basis of such prosodic information, the automatic part of the grammatical coding can be extended and improved. We believe that the incorporation into parsing rules of prosodic and contextual markers, which clearly aid human processing, should be an important and interesting research activity.

The infinitive rule simply states that an unknown word immediately preceded by the infinitive marker *to* (which has a special tag in our system) is probably a verb base.

Finally, the graphic pattern-matching rules predict that words beginning with upper-case letters are proper names, and that words ending in *-ing*, *-ly*, *-ed*, *-est*, *-y* or *-s*

(tried in that order) are *-ing* forms of verbs, adverbs, past tenses of verbs, superlatives of adjectives, adjectives and plural nouns, respectively.

Program performance

A preliminary tagging of some of the conversations in our corpus (subgroup A) was done in the summer of 1980 by two research assistants on the project and two postgraduate students of English. After an initial debugging phase, the tagging program ran quite smoothly, producing 80 per cent correctly tagged words, at a rough estimate. The tags were in fact correct so frequently that it might have been better to let the linguist okay them all at once by letting the program display all the predicted tags belonging to a given tone unit simultaneously, rather than one by one. With two man-months' work, 19 texts, totalling about 95,000 running words, were tagged more or less completely. Fig 5 gives an example of a text tagged for word-classes (pp 56-62).

Tag concordance

During the tagging sessions there arose numerous problems of linguistic interpretation which could hardly have been solved 'in real time' by the linguist sitting at the terminal. In other cases, the availability of a tagged direct-access concordance would have provided the linguist with a valuable aid to making consistent decisions. Ideally, the tagger should have a concordance of previously tagged texts available, enabling the linguist to see not only the current word token in its context, but also all other occurrences of it, including previously assigned tags. The cost of storing such a concordance for direct access is, however, prohibitive. Instead a concordance of the tagged texts has been generated and stored on magnetic tape. It has been sorted

alphabetically by (tagged) key-word in context and printed on paper for proofreading (see Figure 7). Corrections will be made to the tags in the concordance, using the standard text editor CTS. The corrected tags will then be transferred to the tagged texts, to be used for further phrase tagging. An alternative version of the tagged concordance has been sorted by tag, providing a systematic overview of the tagging system.

4.2 SYNTACTIC TAGGING

Our main interest in the use of the corpus is to arrive at a description of the grammar (and related aspects) of spoken English. As indicated on p 51, the word-class and syntactic-phrase tagging programs are interdependent in so far as the output of the former is used as input in the latter and will most likely be modified in the light of the results obtained with the syntactic tagging program.

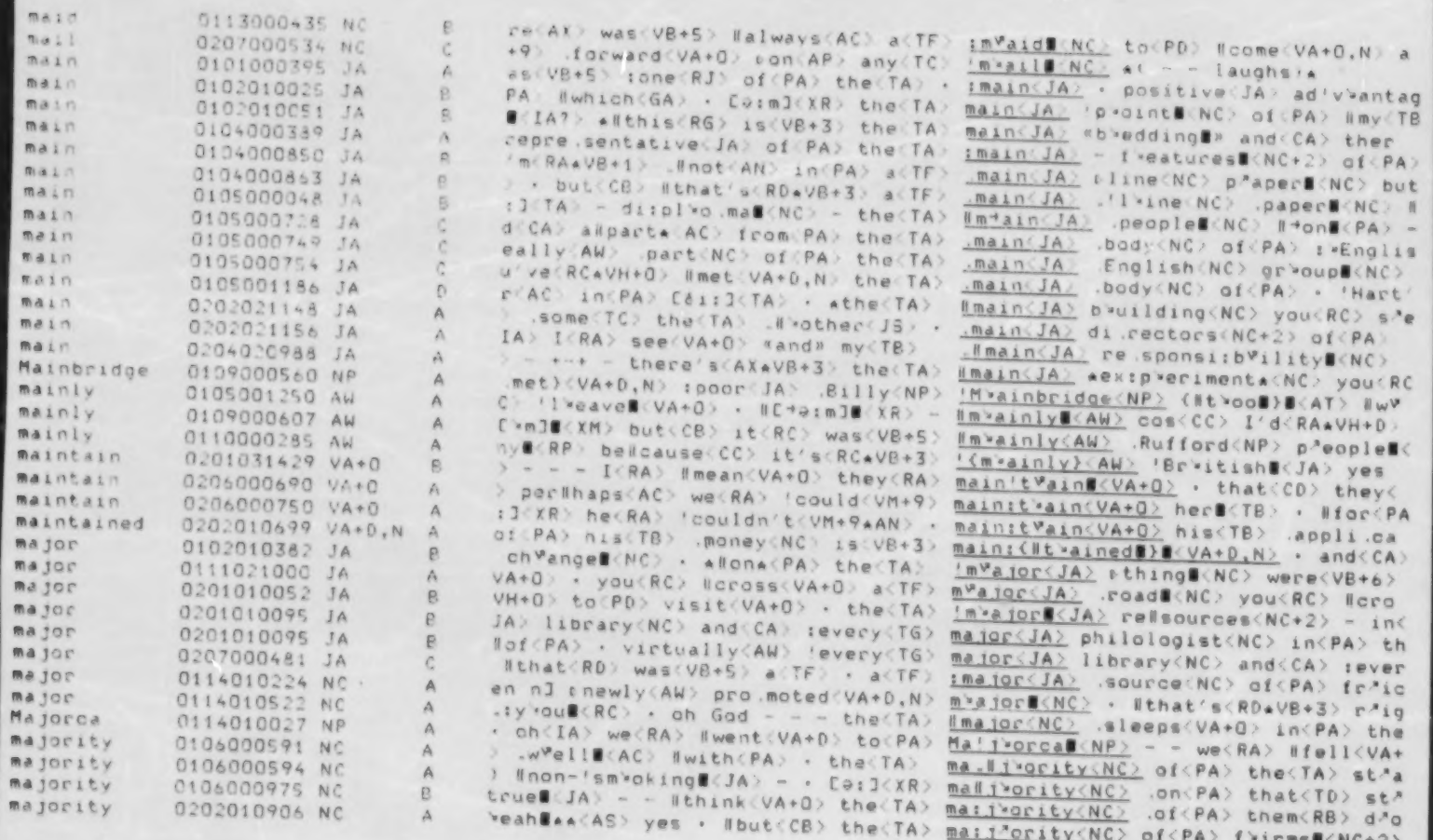
The first problem in writing the algorithms for syntactic analysis is to choose the linguistic unit to be analysed. We decided on the tone unit.

The unit of syntactic analysis

Making the tone unit the basis of our syntactic analysis means that we will break with regular practice which is to start with the sentence. Our reasons for starting with the tone unit are the following.

The material has already been analysed in terms of tone units, whereas the syntactic analysis still remains to be done. By and large, the prosodic analysis has been straightforward with only occasional disagreement among the analysts as to the boundaries of tone units. In particular, the analysis of spoken language in terms of sentences is difficult; some would even say irrelevant (see Svartvik at press). Hence

FIGURE 7. CONCORDANCE SORTED ALPHABETICALLY BY KEY-WORD IN CONTEXT



maid	0113000435	NC	B	re<AX> was<VB+S> ‖always<AC> a<TF>	:mᵛaid■<NC> to<PD> ‖come<VA+D,N> a
mail	0207000534	NC	C	+9) .forward<VA+D> ₵on<AP> any<TC>	'mᵛail■<NC> ▲(- - laughs'▲
main	0101000395	JA	A	as<VB+S> :one<RJ> of<PA> the<TA> ·	:main<JA> · positive<JA> ad'vᵛantag
main	0102010025	JA	B	PA ‖which<GA> · [ə:m]<XR> the<TA>	main<JA> 'pᵛoint■<NC> of<PA> ‖my<TB
main	0102010051	JA	B	■<IA?> ▲‖this<RG> is<VB+3> the<TA>	main<JA> «bᵛedding■» and<CA> ther
main	0104000389	JA	A	cepre.sentative<JA> of<PA> the<TA>	:main<JA> - fᵛeatures■<NC+2> of<PA>
main	0104000850	JA	B	'm<RA▲VB+1> .‖not<AN> in<PA> a<TF>	.main<JA> ₵line<NC> pᵛaper■<NC> but
main	0104000863	JA	B	· but<CB> ‖that's<RD▲VB+3> a<TF>	.main<JA> .'lᵛine<NC> .paper■<NC> ‖
main	0105000048	JA	B	:]<TA> - di:plᵛo.ma■<NC> - the<TA>	‖mᵛain<JA> .people■<NC> ‖ᵛon■<PA> -
main	0105000728	JA	C	d<CA> a‖part▲<AC> from<PA> the<TA>	.main<JA> .body<NC> of<PA> :ᵛEnglis
main	0105000749	JA	C	eally<AW> .part<NC> of<PA> the<TA>	.main<JA> .English<NC> grᵛoup■<NC>
main	0105000754	JA	C	u've<RC▲VH+0> ‖met<VA+D,N> the<TA>	.main<JA> .body<NC> of<PA> · 'Hart'
main	0105001186	JA	D	r<AC> in<PA> [ði:]<TA> · ▲the<TA>	‖main<JA> bᵛuilding<NC> you<RC> sᵛe
main	0202021148	JA	A	.some<TC> the<TA> .‖ᵛother<JS> ·	.main<JA> di.rectors<NC+2> of<PA>
main	0202021156	JA	A	IA> I<RA> see<VA+0> «and» my<TB>	.‖main<JA> re.sponsi:bᵛility■<NC>
main	0204020988	JA	A	- +-+ - there's<AX▲VB+3> the<TA>	‖main<JA> ▲ex:pᵛeriment▲<NC> you<RC
Mainbridge	0109000560	NP	A	.met)<VA+D,N> :poor<JA> .Billy<NP>	'Mᵛainbridge<NP> (‖tᵛoo■)■<AT> ‖wᵛ
mainly	0105001250	AW	A	C 'lᵛeave■<VA+0> · ‖[ᵛə:m]■<XR> -	‖mᵛainly■<AW> cos<CC> I'd<RA▲VH+D>
mainly	0109000607	AW	A	[ᵛm]■<XM> but<CB> it<RC> was<VB+5>	‖mᵛainly<AW> .Rufford<NP> pᵛeople■<
mainly	0110000285	AW	A	ny■<RP> be‖cause<CC> it's<RC▲VB+3>	'(mᵛainly)<AW> 'Brᵛitish■<JA> yes
maintain	0201031429	VA+0	B	> - - - I<RA> ‖mean<VA+0> they<RA>	main't▽ain■<VA+0> · that<CD> they<
maintain	0206000690	VA+0	A	> per‖haps<AC> we<RA> 'could<VM+9>	main:tᵛain<VA+0> her■<TB> · ‖for<PA
maintain	0206000750	VA+0	A	:]<XR> he<RA> 'couldn't<VM+9▲AN> ·	main:tᵛain<VA+0> his<TB> .appli.ca
maintained	0202010699	VA+D,N	A	of<PA> his<TB> .money<NC> is<VB+3>	main:(‖tᵛained■)■<VA+D,N> · and<CA>
major	0102010382	JA	B	ch▽ange■<NC> · ▲‖on▲<PA> the<TA>	'mᵛajor<JA> ₵thing■<NC> were<VB+6>
major	0111021000	JA	A	VA+0> · you<RC> ‖cross<VA+0> a<TF>	mᵛajor<JA> .road■<NC> you<RC> ‖cro
major	0201010052	JA	B	VH+0> to<PD> visit<VA+0> · the<TA>	'mᵛajor■<JA> re‖sources<NC+2> - in<
major	0201010095	JA	B	JA> library<NC> and<CA> :every<TG>	major<JA> philologist<NC> in<PA> th
major	0201010095	JA	B	‖of<PA> · virtually<AW> 'every<TG>	major<JA> library<NC> and<CA> :ever
major	0207000481	JA	C	‖that<RD> was<VB+5> a<TF> · a<TF>	:major<JA> .source<NC> of<PA> frᵛic
major	0114010224	NC	A	en n] ₵newly<AW> pro.moted<VA+D,N>	mᵛajor■<NC> · ‖that's<RD▲VB+3> rᵛig
major	0114010522	NC	A	.₵yᵛou■<RC> · oh God - - - the<TA>	‖major<NC> .sleeps<VA+0> in<PA> the
Majorca	0114010027	NP	A	· oh<IA> we<RA> ‖went<VA+D> to<PA>	Ma'jᵛorca■<NP> - - we<RA> ‖fell<VA+
majority	0106000591	NC	A	.wᵛell■<AC> ‖with<PA> · the<TA>	ma.‖jᵛority<NC> of<PA> the<TA> stᵛa
majority	0106000594	NC	A	) ‖non-'smᵛoking■<JA> - · [ə:]<XR>	ma‖jᵛority<NC> .on<PA> that<TD> stᵛ
majority	0106000975	NC	B	true■<JA> - - ‖think<VA+0> the<TA>	ma:jᵛority<NC> .of<PA> them<RB> dᵛo
majority	0202010906	NC	A	ᵛeah■▲▲<AS> yes · ‖but<CB> the<TA>	ma:jᵛority<NC> of<PA> fᵛirms■<NC+2>

it seems reasonable to start the analysis of our material by looking at the prosodic tone unit and relating its structure to grammatical phrases, clauses and sentences.

It is, furthermore, assumed that the division of stretches of language into tone units is not arbitrary but has linguistic significance. However, there is no agreement on the nature of the relation of prosody to other domains of linguistic structure. Halliday (1967) believes that intonation patterns 'can be systematised into a formal grammatical statement'. Similarly, Brazil (1975) argues that the full significance of intonation can in fact be derived from choices associated with each separate tone unit. On the other hand, Bolinger (1972) takes the view that 'the distribution of sentence accents is not determined by syntactic structure but by semantic and emotional highlighting. Syntax is relevant indirectly in that some structures are more likely to be highlighted than others. But a description along those lines can only be in statistical terms'. From a psycholinguistic point of view, Laver (1970) maintains that 'there is good evidence for believing that ... the tone group is handled in the central nervous system as a unitary behavioural act and the neural correlates of the separate elements are assembled ... before the performance of the utterance begins'. Similarly, Weiner and Goodenough argue that 'the sentence may not be a practical or even "proper" unit for understanding conversational processes' (1977:213). In several recent studies, perceptual strategies are stated in terms of syntactic phrases being processed in portions:

> It is not necessary, however, to assume that these smaller phrases are ever combined into larger syntactic units; they might very well be semantically interpreted immediately, and the pointers between them could have an entirely semantic character. Evidence for the assumption that the smaller syntactic units are never assembled into larger ones is the fact that most spontaneous speech is syntactically quite incoherent from the point of view of written language. Yet we hardly ever notice this until

> we have the opportunity to see speech which has been transcribed verbatim. Thus we seem to attend to syntactic form only in rather short chunks of speech when listening. (Welin 1979:45)

In terms of computer parsing it is in fact an advantage to have tone units as boundaries for the search process. Clause endings are difficult to identify; 'there does not seem to be a reliable indicator of clause ending other than punctuation' (Ross 1973:96). Consequently, in his GUESS algorithm, Ross takes 'only ten or so words' at a time.

Description of the phrase analysis program

The present version of the phrase analysis program analyses input tone units as sequences of phrases, defined within the program as certain word class patterns. For instance, an adjective phrase is defined as a sequence of adjectives, optionally preceded by an adverb phrase, which, in turn, is defined as one or two adjacent adverbs (except certain special adverb types).

The text to be analysed is input tone unit by tone unit. Units uttered by 'non-surreptitious speakers' are immediately skipped. Tone units that contain indistinct passages are also excluded from the analysis.

The tagged words of the remaining tone units are stored sequentially in an array. Contractions are split up into their grammatical components. For example, *don't* will be represented by two tags, one of which corresponds to *do* and the other to *not* (*n't*).

In the main phase of the program, to be described below, the tone unit is scanned for phrases.

When all tone units have been processed, the program prints a statistical summary, showing how tone units are made up of phrases as well as the internal structure of the component phrases. In the future, when more knowledge has been gained about the structure of single tone units, we plan to extend the analysis also over tone unit boundaries.

In the main phase of the program, the text is scanned sequentially from right to left in six passes. In each of the first five passes all instances of one of the major phrase types are constructed from the input. Thus, verb phrases, adverb phrases, adjective phrases, noun phrases, and prepositional phrases are found in that order. In the sixth pass, prepositional phrases are added in a cyclical fashion, as postmodifiers to immediately preceding noun phrases. The ordering of the passes reflects to some extent the progression of the linguistic work done in specifying the phrase patterns, but it is also intrinsic. Output from an early pass may be required as input in a later pass. For example, adverb phrases may be constituents of adjective phrases, which, in turn, may be parts of noun phrases. To be sure, the same effect could be achieved by other means than scanning the input in a certain order. For example, Kay (1977) has described an elegant synchronizing technique using 'waiting lists', whereby lower-order phrases can be added as constituents to higher-order phrases, even if they have not yet been found when the scan for the higher-order phrase is initiated. In our case, the multiple pass approach has enabled us to design the program incrementally and, since the tone units are usually quite short (about 5 words), it probably does not affect performance very much.

The right to left search order within the phrases is considered an important part of the program. It has a certain psychological plausibility as a model of speech perception. Typically, the rightmost word in a phrase is a semantically heavy and prosodically prominent word, which may well be a natural point of departure for the interpretation of preceding, less distinct speech segments, for example:

Verb phrase:	may have been WORKING
Adverb phrase:	most REMARKABLY
Adjective phrase:	most remarkably TALENTED
Noun phrase:	a most remarkably talented PERSON
Prepositional phrase:	in your OFFICE

A common feature of the search procedure of our phrase structure algorithms is that, in most cases, it is based not on the lexical item but on its word class tag and on the sequential order of the letters and numerals of the tag. Thus 'V' in first position of the tag denotes 'verb' in general (and can thus be used for locating any verb in a tone unit); 'M', 'H', 'B', 'D', and 'A' in second position of the tag denote verb subclass (modal, *have*, *be*, *do*, main, respectively). '+Ø', '+3', etc denote morphological form (base form and 3rd person singular, respectively). Thus, together with the letter in second position, the tag can be used for locating verb combinations of two or more 'telescoped' verbs: *may* (VM+8) *come* (VA+Ø) = modal verb phrase (VPH: modal); *might* (VM+9) *have* (VH+Ø) *been* (VB+N) *coming* (VA+G) = verb phrase containing modal, perfect, progressive components (VPH: modal, perf, prog), etc. Combinations of verbs with auxiliary-like qualities, eg *is going to go*, are tagged (VPH: quasi); catenatives, eg *try to get*, are tagged (VPH: cat). A flowchart describing the verb phrase pass is included in Figure 8 (pp 76-81).

Similarly, in a noun phrase like

this most remarkably talented person

the rightmost word *person* is first identified as the head. With the stated search order of grammatical phrases, this noun phrase example will then be analysed roughly as follows:

APH = most remarkably
JPH = APH (most remarkably) talented
NPH = this JPH (APH (most remarkably) talented) person

Since the algorithm for prepositional phrases is applied after the algorithm for noun phrases, the analysis of a prepositional phrase like

for this most remarkably talented person

will be based on the previous noun phrase tag.

Thus an essential feature of the algorithms for phrase tagging is that they are ordered and cyclical. Throughout the tagging process involving phrases (and higher units),

FIGURE 8. FLOWCHART FOR VERB PHRASE ANALYSIS

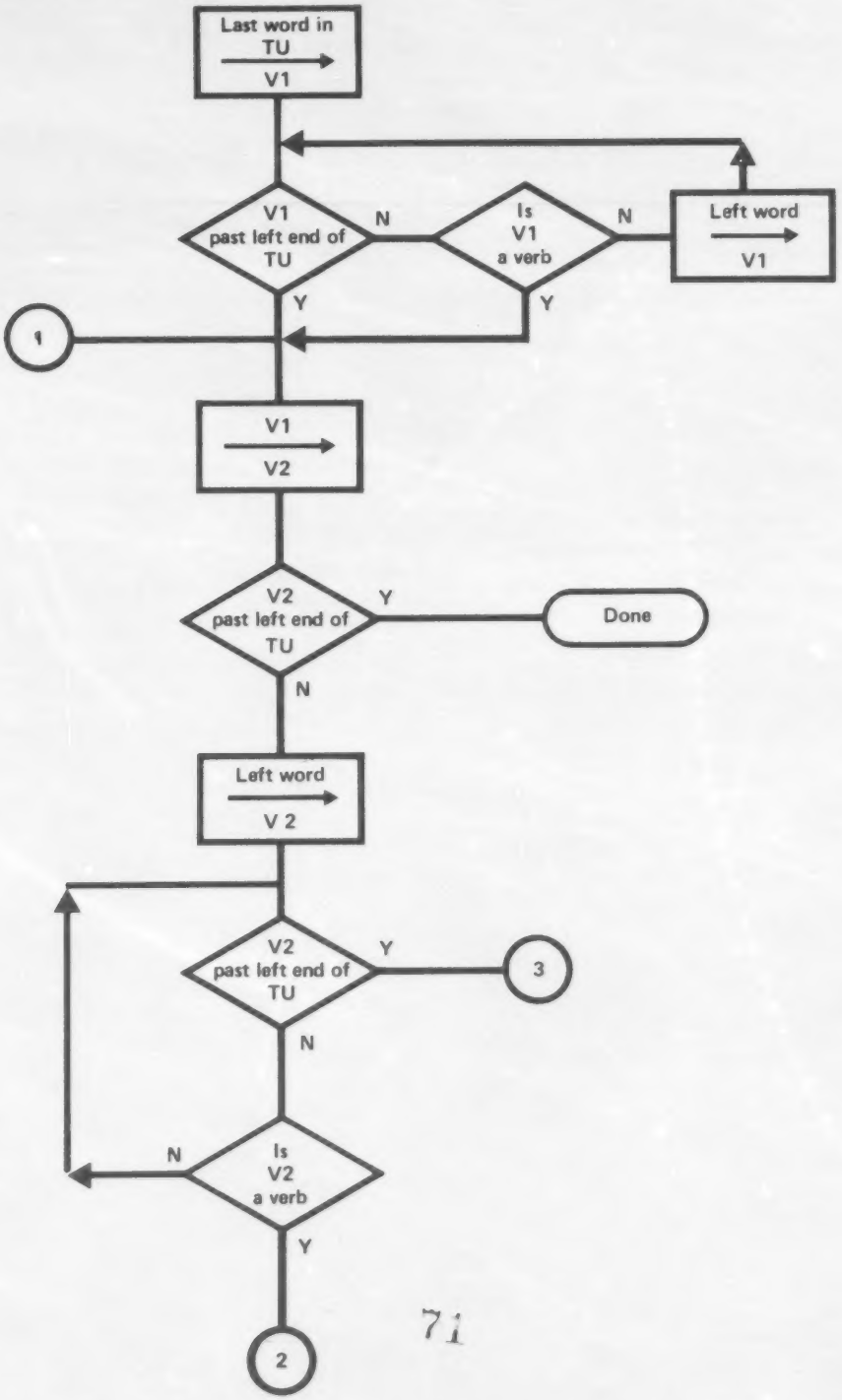


FIGURE 8 CONTINUED

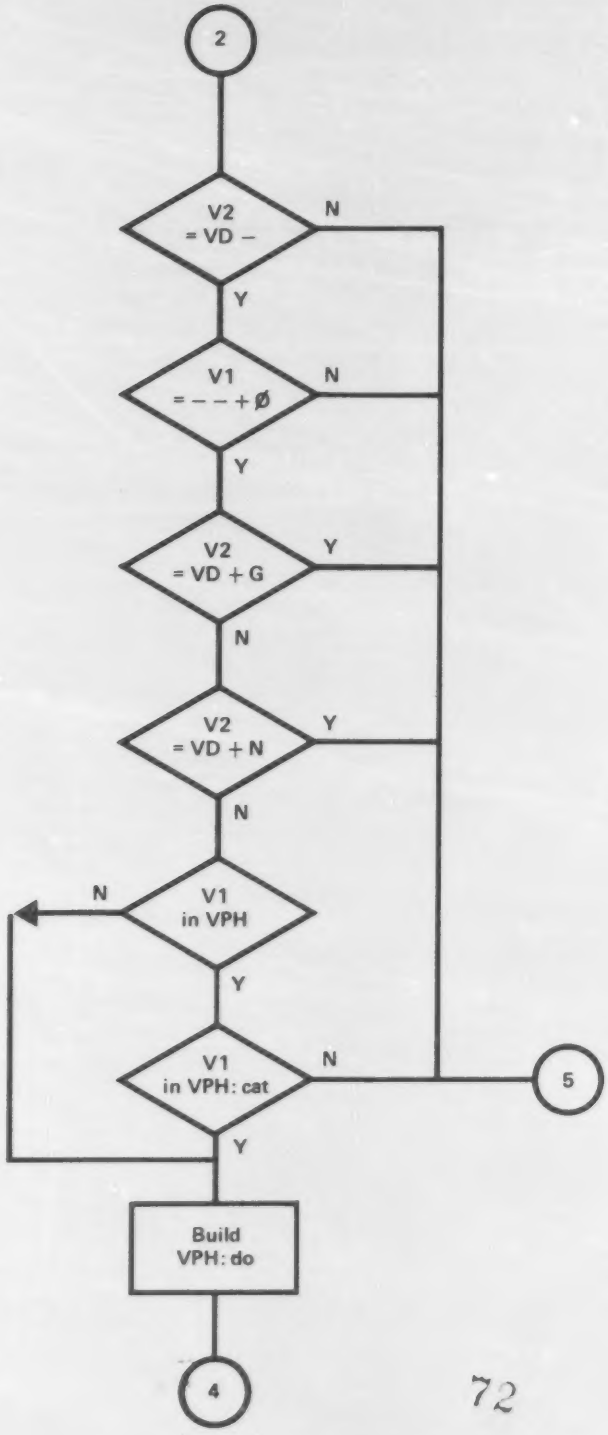


FIGURE 8 CONTINUED

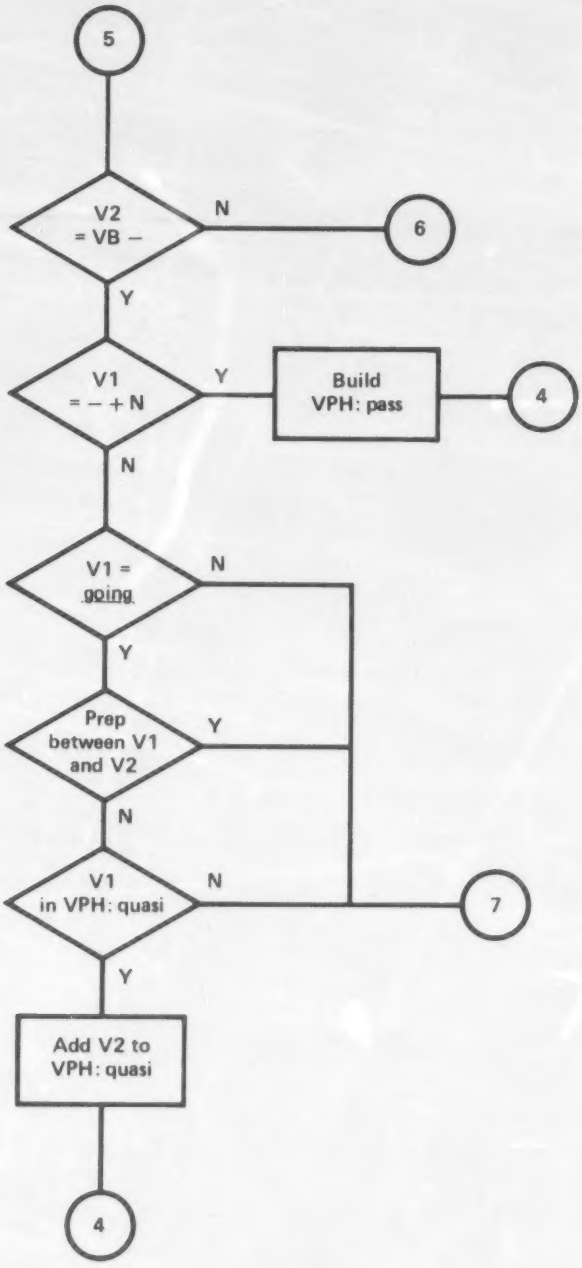


FIGURE 8 CONTINUED

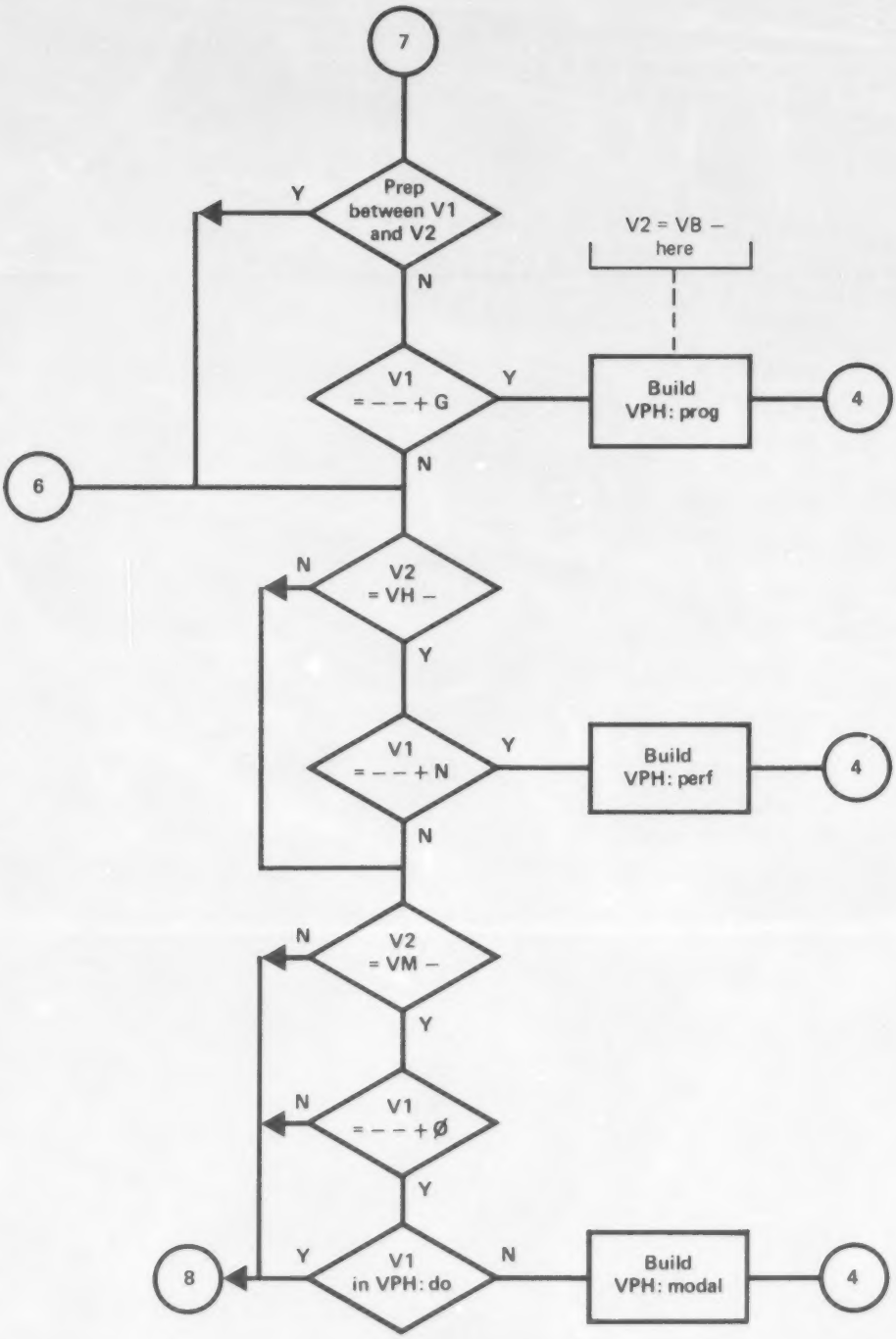


FIGURE 8 CONTINUED

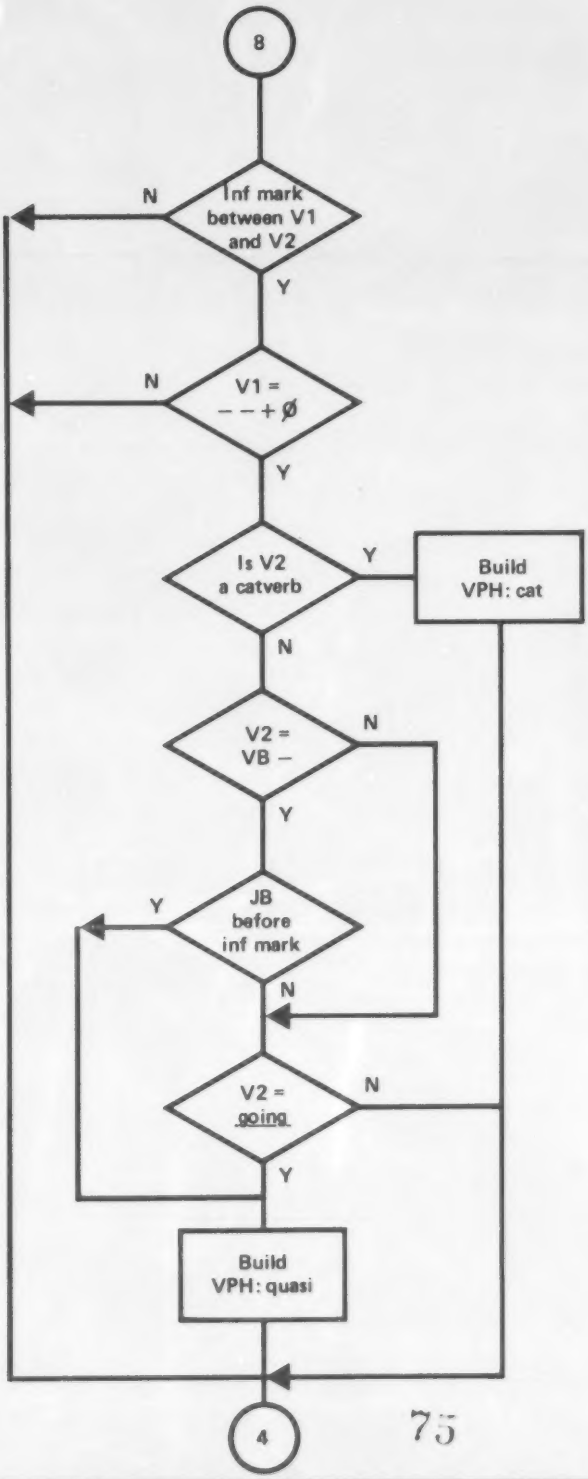
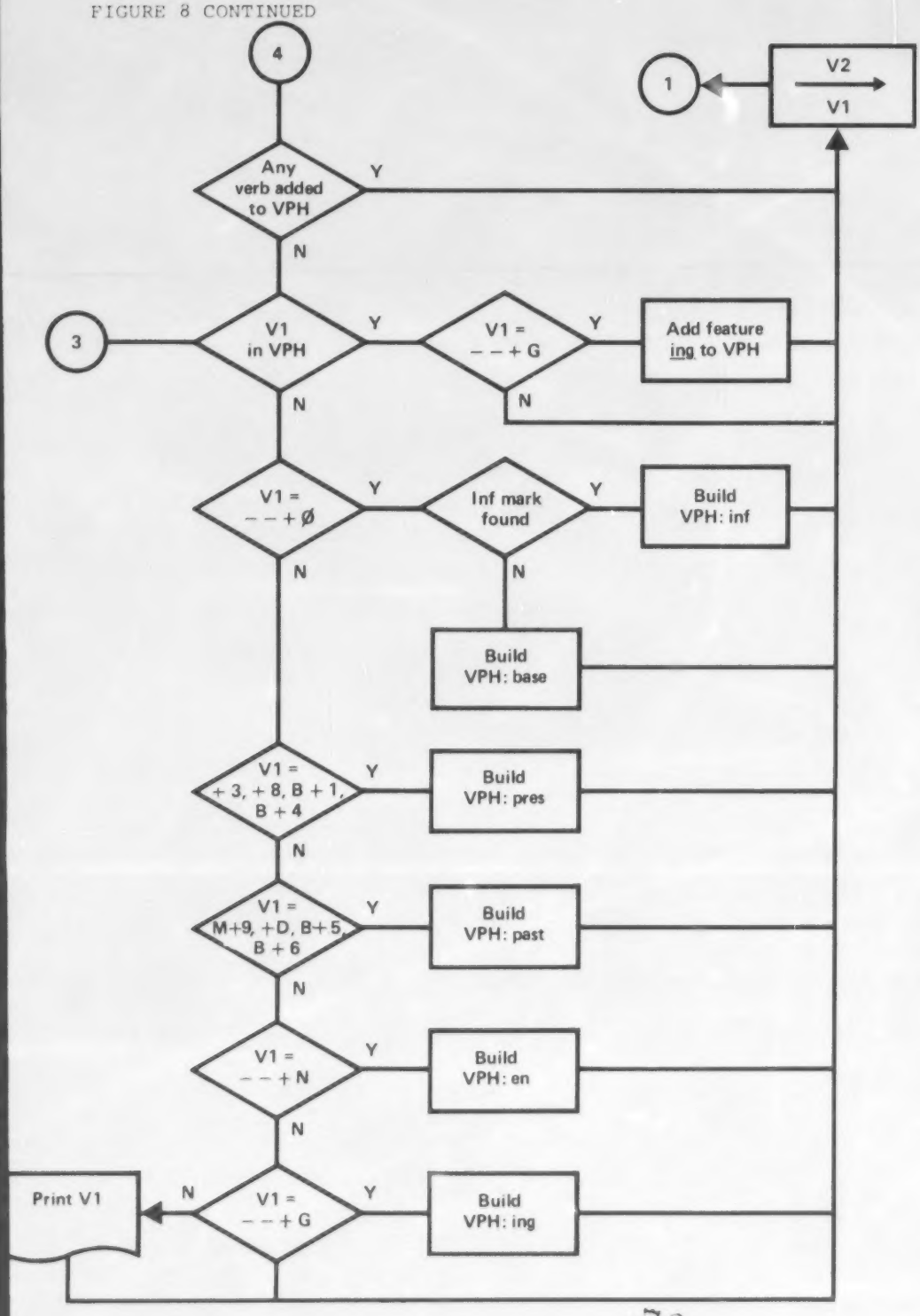


FIGURE 8 CONTINUED



the linguist may be working interactively at the screen, checking and correcting the results of the automatic phrase tagging as well as adding tags where no algorithms exist. The system has been designed to be as open as possible and allow the linguist, in the course of the analysis, to modify old rules in the light of previous results and to introduce new rules when such are called for. Figure 9 shows some examples of output from the verb phrase tagging program.

In its present state, the phrase analysis should be seen basically as a research tool for getting information about the grammatical structure of the prosodically defined tone unit. The present analysis based on grammatical rules incorporated into the program may then be modified to one based on recognition of patterns input as parameters.

Since very little seems to be known at present about the over-all phrase structure of tone units (ie what types of phrases may be included and in what order), no such information is used in the syntactic analysis. In other words, it is essentially bottom-up. In a truly integrated tagging program syntactic expectations based on knowledge about tone unit structure should be used top-down to guide tagging on the word-class level. Extending the program to provide for such feedback is a tempting task, which would lead us to problems closely related to those encountered in automatic speech understanding.

When all phrases in the tone unit have been found, the program prints a list of the maximal ones (ie phrases not included as constituents in higher phrases) from left to right, and indicates whether the analysis is complete, ie whether all words in the tone unit have been assigned to one or more phrases. At present interjections, conjunctions, and certain adverbs are not included in any phrase type. Work is in progress to include such words, too, as special elements on a higher level of structural description of the tone unit, where phrases are to be represented by their (predicted) functions (subject, object etc).

Tagging of clause elements

The grammatical structure of tone units displays great variety. Here are some examples (without prosody) from the illustration in Figure 4, pp 54-55:

(1) [mhm] (TU 562)
(2) yes (TU 549)
(3) the tenth (TU 550)
(4) on Wednesday (TU 571)
(5) and this is to give vivas there (TU 552)
(6) to establish the standards (TU 555)
(7) coming back on the twenty-ninth (TU 551)
(8) have you met our man Yoolet yet (TU 557)
(9) if you could see him in your office sometime (TU 575)
(10) cos he's going to have a lot of responsibility on himself (TU 581)

From a grammatical point of view, these tone units represent two levels: phrase level (examples 1-4) and clause level (examples 5-10). In the case of the latter, we have extended the analysis to include tagging of clause elements, eg (9):

LEXICAL WORDS	PHRASE TAG	CLAUSE ELEMENT TAG
if		sub
you	NPH	S
could see	VPH	V
him	NPH	O
in your office	PPH	A
sometime	APH	A

FIGURE 9. OUTPUT FROM THE VERB PHRASE TAGGING PROGRAM (Text S.1.1, see Figure 5)

```
TU 546  VPH:  prog
           I'm<VB+1> #going<VA+G>

TU 551  VPH:  ing
           #coming<VA+G>

TU 552  VPH:  pres
           is<VB+3>
        VPH:  inf
           to<PD> give<VA+0>

TU 554  VPH:  base
           kn^ow#<VA+0>

TU 555  VPH:  inf
           #to<PD> establish<VA+0>

TU 557  VPH:  perf
           #have<VH+0> ... !m~et<VA+N>

TU 563  VPH:  past
           #knew<VA+0>
        VPH:  prog
           was<VB+5> c~oming#<VA+G>

YU 564  VPH:  perf
           I've<VH+0> #heard<VA+N>
        VPH:  base
           m^ention<VA+0>

TU 567  VPH:  perf
           he's<VH+3> finished<VA+N>
```

```
TU 568   VPH:  base
            ‡kn↗ow■<VA+0>

TU 569   VPH:  modal prog
            ‡you'll<VM+8> be<VB+0> s↘eeing<VA+G>

TU 575   VPH:  modal
            could<VM+9> ‡see<VA+0>

TU 576   VPH:  modal
            I'd<VM+9> :like<VA+0>
         VPH:  inf
            to<PD> have<VH+0>

TU 579   VPH:  cat
            ‡try<VA+0> to<PD> get<VA+0>

TU 580   VPH:  quasi
            he's<VB+3> going<VA+G> to<PD>
             go<VA+0>

TU 583   VPH:  base
            ‡think<VA+0>
         VPH:  prog
            they're★<VB+4> :grooming<VA+G>

TU 585   VPH:  past
            ‡heard<VA+0>
         VPH:  en
            :mentioned<VA+N>
```

The rewrite rules from phrase tags to clause element tags are based on type and relative position of phrase tags. Clearly, even if some such rules can be made to work, the analysis of clause element structure will be limited for a number of reasons. One has already been mentioned: no clause element relations may exist within the tone unit, as in examples 1-4; or the clause elements are only partly to be found in the same tone unit, as in examples 6 and 7. In such cases the analysis will subsequently have to be extended to adjacent tone units.

This patterning is related to a number of factors - prosodic, lexical, grammatical, speaker relation, topic, individual speaker behaviour, and so on. It would seem that the whole problem of segmentation in spoken discourse, which is as complex as it is interesting, stands a better chance of being studied successfully with the help of the techniques outlined here. We plan to pursue this research (for a more detailed account, see Svartvik & Eeg-Olofsson 1982.

4.3 NEGATION

Gunnel Tottie, University of Lund, used the London-Lund Corpus for a paper read at the Åbo Symposium on Impromptu Speech in November, 1981. (G. Tottie & C. Paradis, at press, 'From function to structure. Some pragmatic determinants of syntactic frequencies in impromptu speech', N.E. Enkvist (ed), *Impromptu speech: a symposium*, Åbo Akademi).

Gunnel Tottie is currently using the corpus for an investigation of the variation between synthetic and analytic negation (as in *He saw nothing/He did not see anything*) in conversation vs written English.

Spoken English texts were also used for a study called 'Affixal and non-affixal negation' published in *Studia Linguistica* 34:2, 1980.

The London-Lund Corpus has further been used for term papers at undergraduate level on negative affixes, the pragmatics of negation, modal auxiliaries, verbs denoting mental processes, the use of personal pronouns with modals and 'mental verbs' and a comparative study of the use of adverbials in conversation and in written language.

The following three studies are PhD theses in progress at the University of Lund.

4.4 TURN-TAKING AND INTERRUPTION

Bengt Oreström is working on 'Turn-taking and interruption in face-to-face conversation'. Taking as the point of departure the interactional function of language as manifested in the linguistic behaviour of speaker and listener in face-to-face communication, this study focusses on the turn-taking aspect of dyadic (two-person) conversation. Turn-taking, or the exchange of speaking-turns, is a fundamental concept in conversation and indeed a defining criterion of it; without turn-taking no conversation.

The turn-taking process is studied on the basis of the detailed analysis of 10 conversations in the corpus. Attention is centred on those linguistic features that are correlated with change of speaker and thus operate as signals that guide the listener when he wants to take the floor. Length of speaking-turns is measured by the number of words uttered as well as listener reactions, ie those brief, spontaneous emissions from the listener merely expressing agreement, acceptance and understanding. They are uttered without the speaker intending to take the floor and therefore serve as implicit 'go on' signals, eg *m*, *mhm*, *yes*, *yeah*, *OK*, *exactly*, *fine*, *that's right*, etc. A separate chapter is devoted to those instances of turn-taking which are of a more abrupt nature and/or involve simultaneous talk. An attempt is made to group such cases both according to their form and their function.

4.5 QUESTIONS AND RESPONSES

The aim of Anna-Brita Stenström's thesis 'Questions and responses in English conversation' is to study spoken interaction as it occurs in question/response (Q/R)-sequences in genuine English discourse. Based on a corpus consisting of twenty-five conversations with two or more speakers, it includes face-to-face as well as telephone talk and, for the sake of comparison, a radio interview and a courtroom examination.

The approach is interactive/pragmatic and utterances are analysed on the basis of what they do in the discourse. Q and R are dealt with as moves in exchanges, the exchange being the minimal sequence of discourse interaction. Since all spoken interaction involves turn-taking, the borderline between Q, on the one hand, and statement, on the other, has proved difficult to draw. 'Questionness' is rather a matter of degree, which can be seen along a scale, varying in force from requests for information to requests for acknowledgement. The definition of Q is therefore very wide and includes all verbal utterances, irrespective of form, which elicit a coherently related verbal R. But although there exists no one-to-one relationship between form and function, it is clear that certain syntactic forms, for example, are closely related to particular Q-types.

Q and R are used differently depending on the speech situation: in a friendly chat, for example, as a means of keeping the conversation going; in a radio interview, to inform and entertain an audience; and in the courtroom, to provide relevant information for the jury. The structure of Q/R-exchanges varies depending on such factors as channel of communication, number of interactants, social relations among the speakers, and topic of conversation.

4.6 THIRD PERSON REFERENCE

The title of Cecilia Thavenius's thesis is 'Reference in English conversation. The pragmatics of third person reference.' It is an attempt at a description and a theory of third person pronominal reference in English conversation. The thesis will consist of two main parts. One is an empirical study of third person pronouns in the London-Lund Corpus. The referential patterns and structures of 2365 pronouns have been studied, classified and fed into the computer. Statistical programs of different kinds have been run on the material to provide the results.

The other part of the thesis is an attempt to set up a pragmatic theory of third person reference in English conversation. This part will describe and discuss endophoric and exophoric reference, and the reference relation patterns that have been termed simplexes (one token of a pronoun referring to a referent) and chains (sequences of two or more coreferential pronouns referring to the same referent and with no intervening coreferential pronoun). The functions of chains will be discussed on several levels: the formal, semantic, pragmatic and discourse levels. The individual pronoun groups IT, THEY, HE, and SHE will then be dealt with. A model of the social contextual variables will also be presented. These variables will be discussed in relation to variations in pronominal use.

4.7 OTHER STUDIES

The London-Lund Corpus has also been used by scholars working outside the University of Lund. In addition to the following recent studies reference should also be made to previous work based on the material in the Survey of English Usage (see references in Quirk & Svartvik 1979).

Elisabeth Sahlin, Uppsala University, used the material for a PhD thesis published in 1979: *Some and any in spoken and written English*. (Acta Universitatis Upsaliensis. Studia Anglistica Upsaliensia 38).

Wolf-Dietrich Bald, Technische Hochschule Aachen, used the material for a study called 'Some functions of *yes* and *no* in conversation', published in *Studies in English linguistics for Randolph Quirk*, ed by S. Greenbaum, G. Leech & J. Svartvik, London, 1979.

Joe Taglicht, The Hebrew University of Jerusalem, is using the material for a study on negation and Brita Fjelkestam-Nilsson, Stockholm University, for a study on tag questions.

5 Other Activities

5.1 LECTURES, CONFERENCES AND SYMPOSIA

During the period of the project members of the research group have lectured on the Survey or connected topics and attended meetings as follows.

Jan Svartvik lectured at the following universities:
Copenhagen, Edinburgh, Helsinki, Rostock, Saarbrücken, Stockholm, Trondheim, Växjö and Åbo Akademi;

lectured at conferences and symposia in

Poznań: The Tenth International Congress of the International Association of University Professors of English;
Washington, D.C.: The International Congress of Fédération Internationale des Langues Vivantes;
Austin, Texas: The Third Conference of Nordic and General Linguistics;
Montreal: The Fifth International Congress of Applied Linguistics;
Södertälje: A symposium organized by Associations Scandinaves de Linguistique Appliquée (AScLA);
Albany, N.Y., The State University of New York: A symposium called "Apollo Agonistes";
Bergen: A Symposium on Grammatical Tagging of English Text Corpora in Machine-Readable Form;
Stockholm: Nobel Symposium on Text Processing;
Bielefeld: Workshops on the Semiotics of Intonation;
Åbo: The Linguistic Study of Impromptu Speech;

lectured to groups of language teachers in various places, including Stockholm, Göteborg, Karlskrona, Solbacka, Avesta, Malmö, Laholm, Kalmar, Trollhättan.

Jan Svartvik and Mats Eeg-Olofsson talked about the project
at a symposium in Lund on data processing and its application with participants from the Christian Albrechts Universität, Kiel;

at a postgraduate seminar in the Department of English, Stockholm;

at a seminar on Computer Corpora in Research and Teaching in Bergen.

Bengt Oreström lectured at
the Departments of English, Linguistics, and Phonetics, Lund University;

Laval University, Quebec;

lectured at conferences and symposia in
Stockholm: Svenska Föreningen för Vetenskapsjournalistik
Montreal: The Fifth International Congress of Applied Linguistics;
Leeds: The Teaching of Spoken English;
Åbo: The Linguistic Study of Impromptu Speech;

attended the following conferences and symposia
Göteborg: The First Swedish Colloquium on Computational Linguistics;
Södertälje: A symposium organized by AScLA;
Copenhagen: De nordiske datalingvistikdage 1979;
Milwaukee: A conference called "Beyond the Sentence";

visited departments of English and Linguistics at
the University of Edinburgh;
the University of Lancaster;
the University of Birmingham;
the Survey of English Usage, University College London;

and visited the Shetland Islands, Scotland, for a linguistic experiment with material from the London-Lund Corpus.

Cecilia Thavenius lectured at
the Departments of English and Phonetics, Lund University;

lectured at conferences and symposia in
Montreal: The Fifth International Congress of Applied Linguistics;
Copenhagen: De nordiske datalingvistikdage 1979;
Åbo: The Linguistic Study of Impromptu Speech;

attended the following conferences and symposia
Göteborg: The First Swedish Colloquium on Computational Linguistics;
Göteborg: Nordiska datalingvistikdagar 1977;
Frostavallen: The Fifth Scandinavian Conference of Linguistics;
Leeds: The Teaching of Spoken English;
Lysebu, Oslo: The First Nordic Conference for English Studies;
Philadelphia: The Tenth New Wave Conference;

and visited departments of English and Linguistics at
the University of Edinburgh;
the University of Lancaster;
the Survey of English Usage, University College London;
Georgetown University, Washington, D.C.;
The Center for Applied Linguistics, Washington, D.C.;
MIT and Harvard University, Cambridge, Mass.

The research group were interviewed by Sveriges Radio for 'Vetandets värld', broadcast on 31 October, 1979.

5.2 SEMINARS AND COLLOQUIA ARRANGED BY THE SURVEY

In September 1978 a seminar was arranged with invited specialists on the computational aspects of the second phase of the project.

In November 1978 there took place a seminar on discourse analysis with participants from the Universities of Copenhagen and Lund.

In February 1979 the Survey hosted the Second Swedish Colloquium on Computational Linguistics at Lund.

In March 1979 the compilers of the Brown University Corpus, Professors W Nelson Francis and Henry Kučera, visited the Survey and gave a seminar with participants from Lund and Copenhagen.

5.3 VISITORS TO THE SURVEY

Raymond Adlam, The British Council, Stockholm
Sture Allén, University of Göteborg
Jens Allwood, University of Göteborg
Wolf-Dietrich Bald, Technische Hochshule, Aachen, FRG
Inger and Bernhard Bierschenk, Lärarhögskolan, Malmö
Lennart Björk, University of Stockholm
David Brazil, University of Birmingham, UK
Benny Brodda, University of Stockholm
Ruth and Kenneth Brown, Canadian Embassy, Stockholm
László Budai, Eger, Hungary
Jens Christensen, University of Copenhagen, Denmark
Pit Corder, University of Edinburgh, UK
Ingemar Dahlstrand, University of Lund
Dick Dawson, University of Salisbury, Zimbabwe
Alvar Ellegård, University of Göteborg
Gunnel Engwall, University of Stockholm
Nils Erik Enkvist, Åbo Akademi, Finland
Claus Faerch, University of Copenhagen, Denmark
Jacek Fisiak, Adam Mickiewicz University, Poznań, Poland
Henni Forchhammer, Handelshögskolan, Copenhagen, Denmark
Charles Foster, Department of Education, Washington, D.C, USA
Nearlene and W Nelson Francis, Brown University, Providence, R.I., USA
Rolf Gavare, University of Göteborg
Eugene Green, Boston University, Mass, USA
Peter Harder, University of Copenhagen, Denmark
Börje Hegbrant, Trelleborg

Helmut Henne, Technische Universität, Braunschweig, FRG
Jesper Hermann, University of Copenhagen, Denmark
Dick Hudson, University College London, UK
Jan Hultgren, University of Stockholm
Robert Ilson, Institute of Education, London, UK
Sven Jacobson, University of Stockholm
Torben Jensen, University of Copenhagen, Denmark
Nils Lennart Johannesson, University of Stockholm
Faith-Ann and Stig Johansson, University of Oslo, Norway
David Jones, University of Stockholm
Bengt Jürgensen, Handelshögskolan, Copenhagen, Denmark
Henry Kučera, Brown University, Providence, R.I., USA
Britt-Marie Kylander, Torsby
Pentti Lappalainen, Alvesta
Geoffrey Leech, University of Lancaster, UK
Rita Lenstrup, University of Copenhagen, Denmark
Magnus Ljung, University of Stockholm
Jonas Löfström, University of Göteborg
Bente Maegaard, University of Copenhagen, Denmark
Annika Mattsson, University of Uppsala
Raven McDavid, University of Chicago, USA
Gunnel Melchers, University of Stockholm
Michael Mepham, Laval University, Quebec, Canada
David Mighetto, University of Göteborg
Keith Morrow, University of Reading, UK
Charles Muller, University of Strasbourg, France
Eugene Nida, United Bible Societies, New York, USA
Naoyuki Okada, Oita University, Japan
Randolph Quirk, University College London, UK
Earl Rand, UCLA, Los Angeles, Calif, USA
Jan Rusiecki, University of Warsaw, Poland
Hanne Ruus, University of Copenhagen, Denmark
David Sankoff, University of Montreal, Canada
Theodore Sellin, US Consulate General, Göteborg
Denys Sharrocks, British Council, Stockholm
John Sinclair, University of Birmingham, UK

Christina Stendahl, University of Göteborg
Anna-Lena Sågvall-Hein, University of Uppsala
Hans Söderlind, University of Uppsala
Mette Cathrine Jahr Sørheim, University of Oslo, Norway
Jan-Gunnar Tingsell, Lärarhögskolan, Mölndal
Tone Tingsgård, University of Uppsala
Peter Trudgill, University of Reading, UK
Janet Whitcut, Longman Group, Harlow, Essex, UK
Kay Wikberg, University of Tromsø, Norway

6 Publications by Survey Staff 1975–1982

1975 (1) Leech, Geoffrey & Jan Svartvik. *A communicative grammar of English*. London: Longman.

(2) Svartvik, Jan & Bengt Törjas. Rhythmic variation in the use of time references. *Style and Text. Studies presented to Nils Erik Enkvist*, ed by H. Ringbom, 416-432. Stockholm: Skriptor.

1976 (3) Oreström, Bengt, Jan Svartvik & Cecilia Thavenius. Manual for terminal input of spoken English material. SSE report. Lund: Lund University.

(4) Svartvik, Jan. Projektet Engelskt talspråk. SSE report. Lund: Lund University.

(5) Thavenius, Cecilia. A select bibliography of studies in spoken English. SSE report. Lund: Lund University.

1977 (6) Oreström, Bengt. Supports in English. SSE report. Lund: Lund University.

(7) Oreström, Bengt. Why "/ði/ book"? SSE report. Lund: Lund University.

(8) Svartvik, Jan & David Wright. The use of *ought* in teenage English. *Acceptability in language*, ed by S. Greenbaum (= *Contributions to the sociology of language* 17, ed by J.A. Fishman), 179-201. The Hague: Mouton.

(9) Svartvik, Jan. A proposal for interactive analysis of spoken English. SSE report. Lund: Lund University.

(10) Thavenius, Cecilia. A select bibliography of studies in spoken English. Rev. ed. SSE report. Lund: Lund University.

1978 (11) Oreström, Bengt. Turn-taking: On speaker-shift in face-to-face conversation. SSE report. Lund: Lund University.

(12) Oreström, Bengt. Avbryt inte! *Forskning och framsteg* 3, 28-30.

(13) Oreström, Bengt & Cecilia Thavenius. Auditory and acoustic analysis: An experiment. SSE report. Lund: Lund University.

(14) Petti, Vincent & Jan Svartvik. A reform in phonetic notation. *Lingua* 1978:2, 73-75, *Moderna Språk* 72:2, 113-116.

(15) Svartvik, Jan. The contribution of Otto Jespersen. *The Nordic languages and modern linguistics 3. Proceedings of the Third Conference of Nordic and General Linguistics*, ed by J. Weinstock, 65-69. Austin, Texas: University of Texas.

(16) Svartvik, Jan. Att undersöka engelskt talspråk. *Lingua* 1978:1, Språkpedagogik 1-3.

(17) Svartvik, Jan & Olof Sager. *Engelsk universitetsgrammatik*. Stockholm: Esselte Studium.

(18) Thavenius, Cecilia. A pilot study of referential 'it' in spoken English. SSE report. Lund: Lund University.

1979 (19) Oreström, Bengt. Rapport från konferensen The Teaching of Spoken English i Leeds, *Lingua* 1979:4.

(20) Quirk, Randolph & Jan Svartvik. A corpus of modern English. SSE report. Lund: Lund University. (Also in *Empirische Textwissenschaft. Aufbau und Auswertung von Text-Corpora*, ed by H. Bergenholtz & B. Schaeder. Königstein: Scriptor.)

(21) Svartvik, Jan & Moira Linnarud (eds). *Kommunikativ kompetens och fackspråk*. Lund: Svenska föreningen för tillämpad språkvetenskap.

(22) Thavenius, Cecilia. Projektet Engelskt talspråk. Nordiske Datalingvistikdage i Khvn 9-10 oktober 1979. Foredrag udg af B. Maegaard. København: Københavns Universitet.

(23) Thavenius, Cecilia. Rapport från AILA-kongressen 1978. *Lingua* 1978:4, Språkpedagogik 16.

(24) Thavenius, Cecilia & Bengt Oreström (eds). Konkordanser: Föredrag från 2:a svenska kollokviet i språklig databehandling i Lund. SSE report. Lund: Lund University.

1980 (25) Oreström, Bengt. Språksyn och samhällssyn. *Lingua* 1980:1.

(26) Oreström, Bengt. Turn length in face-to-face conversation. SSE report. Lund: Lund University.

(27) Svartvik, Jan. Tagging spoken English. *ALVAR. A linguistically varied assortment of readings. Studies presented to Alvar Ellegård on the occasion of his 60th birthday*, ed by J. Allwood & M. Ljung. Stockholm: Department of English.

(28) Svartvik, Jan. Computer-aided grammatical tagging of spoken English. SSE report. Lund: Lund University.

(29) Svartvik, Jan. Well in conversation. *Studies in English linguistics for Randolph Quirk*, ed by S. Greenbaum, G. Leech & J. Svartvik, 167-177. London: Longman.

(30) Svartvik, Jan. Interactive parsing of spoken English. *Proceedings of the 8th International Conference on Computational Linguistics, Tokyo, September 30 - October 4, 1980.*

(31) Svartvik, Jan & Randolph Quirk (eds). *A corpus of English conversation*. Lund Studies in English 56. Lund: Gleerup/Liber.

(32) Thavenius, Cecilia. Review of Henne, H. & H. Rehbock, *Einführung in die Gesprächsanalyse*, *Studia Linguistica* 1980:1.

1981 (33) Oreström, Bengt. Att undersöka samtal. *Språkteorins relevans för språkundervisningen*, 63-68, ed by J. Allwood. (Rapport från ett symposium för språkforskare och språklärare.) Göteborg: Studieförlaget.

(34) Thavenius, Cecilia. Referential 'it' in Spoken English. *Actes du V^e Congrès de l'Association Internationale de Linguistique Appliquée Montreal août 1978*, éd par J.-G. Savard & L. Laforge. Québec.

1982 (35) Svartvik, Jan & Mats Eeg-Olofsson. Tagging the London-Lund corpus of spoken English. *Computer corpora in English language research*, ed by S. Johansson. Bergen: Norwegian Computing Center for the Humanities.

At press (36) Oreström, Bengt. When is it my turn to speak? *Impromptu speech: a symposium*, ed by N.E. Enkvist. Åbo: Åbo Akademi.

(37) Svartvik, Jan. Information processing in speech. Preprints for the Nobel Symposium on Text Processing, Stockholm, August 11-15 1980, ed by S. Allén.

(38) Svartvik, Jan. On speaking terms. Inträdesföredrag, Kungl Vitterhets Historie och Antikvitets Akademien, 1 december 1981. To be published in *Akademiens Årsbok*.

(39) Svartvik, Jan. The segmentation of impromptu speech. *Impromptu speech: a symposium*, ed by N.E. Enkvist. Åbo Akademi.

(40) Thavenius, Cecilia. Exophora in English conversation. A study of third person pronominal reference. *Impromptu speech: a symposium*, ed by N.E. Enkvist. Åbo: Åbo Akademi.

References

Bolinger, D. Accent is predictable (if you're a mindreader).
1972 *Language* 48:3, 633-644.

Brazil, D. *Discourse intonation*. Discourse analysis mono-
1975 graphs 1. Birmingham: Department of English.

Crystal, D. *Prosodic systems and intonation in English*.
1969 Cambridge: Cambridge University Press.

Crystal, D. & D. Davy. *Advanced conversational English*.
1975 London: Longman.

Crystal, D. & R. Quirk. *Systems of prosodic and para-*
1964 *linguistic features in English*. Janua Linguarum, Series Minor 39. The Hague: Mouton.

Francis, W.N. Problems of assembling and computerizing large
1979 corpora. *Empirische Textwissenschaft. Aufbau und Auswertung von Text-Corpora*, ed by H. Bergenholtz & B. Schaeder, 110-123. Königstein: Scriptor.

Francis, W.N. A tagged corpus - problems and prospects.
1980 *Studies in English linguistics for Randolph Quirk*, ed by S. Greenbaum, G. Leech & J. Svartvik, 192-209. London: Longman.

Francis, W.N. & H. Kučera. *Manual of information to accompany*
1964 *a standard sample of present-day edited American English, for use with digital computers*. Revised eds 1971 and 1979. Providence: Department of Linguistics, Brown University.

Francis, W.N. & H. Kučera. *Frequency analysis of English usage:*
At press *lexicon and grammar*. Boston: Houghton Mifflin.

Halliday, M.A.K. *Intonation and grammar in British English*.
1967 The Hague: Mouton.

Hofland, K. & S. Johansson. *Word frequencies in British and*
1982 *American English*. Bergen: Norwegian Computing Centre for the Humanities.

Johansson, S. (ed). *Computer corpora in English language*
1982 *research*. Bergen: Norwegian Computing Centre for the Humanities.

Johansson, S., G. Leech & H. Goodluck. *Manual of information*
1978 *to accompany the Lancaster-Oslo/Bergen corpus of British English, for use with digital computers*. Oslo: Department of English, University of Oslo.

Kay, M. Morphological and syntactic analysis. *Linguistic struc-*
1977 *ture processing*. Ed by Antonio Zampolli, 131-234. Amsterdam: North-Holland.

Kučera, H. & W.N. Francis. *Computational analysis of present-*
1967 *day American English*. Providence: Brown University Press.

Laver, J. The production of speech. *New horizons in linguistics*,
1970 ed by J. Lyons, 53-77. Harmondsworth: Penguin.

Longman dictionary of contemporary English (LDOCE), ed by
1978 P. Procter. London: Longman.

Norén, S. Konkordanser till the London-Lund corpus. Konkor-
1979 danser: Föredrag från 2:a svenska kollokviet i språklig databehandling i Lund, ed by C. Thavenius & B. Oreström. SSE report. Lund: Lund University.

Quirk, R., S. Greenbaum, G. Leech & J. Svartvik. *A grammar of*
1972 *contemporary English*. London: Longman.

Quirk, R. & J. Svartvik. A corpus of modern English. SSE
1979 report. Lund: Lund University. (Also in *Empirische Textwissenschaft. Aufbau und Auswertung von Text-Corpora*, ed by H. Bergenholz & B. Schaeder. Königstein: Scriptor.)

Ross, D. Beyond the concordance: Algorithm for description
1973 of English clauses and phrases. *The computer and literary studies*, ed by A.J. Aitken, R.W. Bailey & N. Hamilton-Smith, 85-99. Edinburgh: University Press.

Svartvik, J. The segmentation of impromptu speech. *Impromptu*
At press *speech: a symposium*, ed by N.E. Enkvist. Åbo Akademi.

Svartvik, J. & M. Eeg-Olofsson. Tagging the London-Lund corpus
1982 of spoken English. *Computer corpora in English language research*, ed by S. Johansson. Bergen: Norwegian Computing Centre for the Humanities.

Svartvik, J. & R. Quirk (eds). *A corpus of English conversation*.
1980 Lund Studies in English 56. Lund: Gleerup/Liber.

Weiner, S.L. & D.R. Goodenough. A move toward a psychology of
1977 conversation. *Discourse processes: Advances in research and theory 1: Discourse production and comprehension*, ed by R.E. Freedle, 213-225. Norwood, New Jersey: Ablex.

Welin, C.W. *Studies in computational text comprehension*. Mono-
1979 graph 5, Institute of Linguistics. University of Stockholm.

Appendix: Word-Class Tags

A *ADVERBS*

AB *wh*-adverbs: *where, why, how ...*
AC Closed class adverbs (other than those specified here otherwise), ie those not derived from adjectives: *slow, early, there, just, far ...*
AE Postpositional adverbs: *ago, enough ...*
AF *more* (all uses)
AG *most* (all uses)
AI *very*
AM *much*
AN *not*
AP Adverbs as particles in phrasal verbs (*to set up*) or nouns (*the washing up*), but prepositions in prepositional verbs, *walk up the street*, = PA
AR *no*: 'no more', 'no good' ...
AS *yes, yeah, yea, aye ...*
AT *too*
AW Open class adverbs, ie those ending in *-ly* and derived from adjectives: *easily ...*
AX *there* as existential adverb (locative *there* = AC)
AZ *so* (all uses)
+R comparative: *earlier* = AC+R
+T superlative: *earliest* = AC+T

C *CONJUNCTIONS*

CA *and*
CB *but*
CC All subordinators except *that*
CD Subordinating *that*: 'She thought that ...'
CE Correlative *as, than*
CR *or*
CS *both, either, neither*

E *PREDETERMINERS*

EC *all*: 'all the time'
ED *half*: 'half a pint'
EE *both*: 'both days'
EF *double*: 'double the income'
EJ *many*: 'many a day'
EK *what, such*: 'what bad luck'
EL *quite, rather*: 'quite a girl'

G *RELATIVE PRONOUNS*

GA *who, which*
GB *whom*
GC *what* (without a nominal head): 'what he says is true'
GD *that*: 'the car that is parked over there'

I *INTERJECTIONS*

IA Interjections 'proper': *oh*, *gosh* ...
Other word-classes used as interjections are coded thus: *fine* JA = I, *God* NP = I, *well* AC = I, *what* RF = I, *thanks* NC+2 = I ...

J *ADJECTIVES*

JA Adjectives other than those specified below
JB 'be *able* to', 'be *about* to'
JE Postposed adjective: 'president elect'
JM *few*, *least*, *less*, *little*, *many*, *more*, *most*, *same*, *several*, *such* } + noun: '(a) few (young) people' ...
own (all uses): 'my own car' ...
JN Adjectives denoting nationality and/or language: *English* ...
JP Adjectives as noun phrase head: 'the dead', 'the accused', 'the same' ...
JQ Ordinals: *first*, *second* ...
JR Cardinals: *one*, *two* ... and pseudocardinals: *fortyish* ...
+R comparative: *finer* = JA+R
+T superlative: 'do one's best' = JP+T

N *NOUNS*

NC Nouns other than those specified below
NN Nationality nouns: *Englishman* ...
NP Proper nouns: *England*, *John* ..., including eg titles before names (*Sir John*) and book titles
NX Nominal abbreviations, acronyms etc: *BBC*

+2 plural: *boy* = NC, *boys* = NC+2
+Z genitive: *boy's* = NC+Z, *boys'* = NC+2+Z

P *PREPOSITIONS*

PA Prepositions other than PD
PD Infinitive marker *to*

R *PRONOUNS OTHER THAN RELATIVE (with non-determiner function)*

RA Personal/subjective: *I*, *he*, ...
RB Personal/objective: *me*, *him* ...
RC Personal unmarked for case: *you*, *it*
RD Demonstrative *that*: 'that's a nice car'

RE Possessive with non-determiner function: *mine* ...
RF Interrogative: *who, whose, which, what* (± *ever*)
RG Demonstrative *this*: 'this is terrible'
RH Demonstrative *these, those*: 'how much are these?'
RJ *One, ones* (all uses except numeral): 'that's a big one'

RM Pronouns in *-body, -one, -thing* (transcribed as one word):

somebody	*someone*	*something*
anybody	*anyone*	*anything*
everybody	*everyone*	*everything*
nobody	*noone*	*nothing*

RN *else*
RP Quantifiers:

all	*either*	*many*	*other(s)*
another	*enough*	*more*	*plenty*
any	*few*	*most*	*several*
both	*less*	*much*	*some*
each	*little*	*neither*	*such*
		none	

RR Reflexives: *myself, ourselves* ...

T DETERMINERS

TA *the*, [dhi(:)]: 'the car' ...
TB *my, no, whose, which, whatever*: 'whose car' ...
TC *some, any, enough*: 'any car' ...
TD *this, that*: 'that car' ...
TE *these, those*: 'those cars' ...
TF *a, an*, [ei], [æn] : 'a car' ...
TG *every, each, either, neither, another*: 'every car' ...
TH *much*: 'much trouble' ...

V VERBS

	MAIN VERBS VA	*BE* VB (all uses)	*DO* VD (all uses)	*HAVE* VH (all uses)	*MODALS* VM
+Ø	*speak* *walk*	*be*	*do*	*have* *'ve*	
+3	*speaks* *walks*	*is* *'s*	*does*	*has* *'s*	
+D	*spoke* *walked*		*did*	*had* *'d*	
+N	*spoken* *walked*	*been*	*done*	*had*	
+G	*speaking* *walking*	*being*	*doing*	*having*	

	MAIN VERBS	*BE*	*DO*	*HAVE*	*MODALS*
	VA	VB	VD	VH	
		(all uses)	(all uses)	(all uses)	
+1		*am*			
		'm			
+4		*are*			
		're			
+5		*was*			
+6		*were*			
+8					*can, may ...*
+9					*could, might ...*

Marginal verbs, eg *need*, are coded thus:

VM+8 She *need* not go

VA+Ø She does not *need* to go

Contractions (indicated by asterisk)

VB+3*AN	*isn't*
VB+4*AN	*aren't*
VB+5*AN	*wasn't*
VB+6*AN	*weren't*
RC*VB+4	*you're*
RA*VB+1	*I'm*
RA*VH+Ø	*I've*
RA*VB+3	*he's* (here)
RA*VH+3	*he's* (done it)
V%+3	*'s* when no way of saying whether it is a *be* or *have* form: '*he's* gone' ...
RA*VH+D	*she'd* (done it)
RA*VM+9	*she'd* (come if she could)
RA*VM+8	*they'll*
V%+7	*ain't*

X MISCELLANEOUS

XA	Metalanguage: 'the word *like*'
XM	/m/, /mhm/
XR	/ə/, /ə:/, /əm/, /ə:m/
XX	Foreign words: *Kaiser, ad hominem*, formulae: *a log ...*, separate letters: *J O H N*
XZ	general 'ragbag' /hm/, /uhuh/ ...

QUERIES

When in doubt about word-class assignment, either bare '?' or 'NC?', 'VA+D?', etc have been used.

LUND STUDIES IN ENGLISH

Founded by *Professor Eilert Ekwall*

1. Bertil Weman, Old English Semantic Analysis and Theory with Special Reference to Verbs Denoting Locomotion. 1933. 187 pp.

2. Hilding Bäck, The Synonyms for "Child", "Boy", "Girl" in Old English. An Etymological-Semasiological Investigation. 1934. XVI+273 pp.

3. Gustaf Fransson, Middle English Surnames of Occupation. With an Excursus on Toponymical Surnames. 1935. 217 pp.

4. G. Hofstrand, The Seege of Troye. A Study in the Intertextual Relations of the Middle English Romance the "Seege or Batayle of Troye". 1936. XV+205 pp.

5. Urban Ohlander, Studies on Coordinate Expressions in Middle English. 1936. 213 pp.

6. Viktor Engblom, On the Origin and Early Development of the Auxiliary Do. 1938. 169 pp.

7. Ivar Dahl, Substantival Inflexion in Early Old English. Vocalic Stems. 1938. XVI+206 pp.

8. Hilmer Ström, Old English Personal Names in Bede's History. An Etymological Investigation. 1939. XLIII+180 pp.

9. Uno Philipson, Political Slang 1750–1850. 1941. XVI+314 pp.

10. Arthur H. King, The Language of Satirized Characters in Poetaster. A Socio-Stylistic Analysis 1579–1602. 1941. XXXIV+258 pp.

11. Mattias T. Löfvenberg, Studies on Middle English Local Surnames. 1942. XLV+225 pp.

12. Johannes Hedberg, The Syncope of the Old English Present Endings. A Dialect Criterion. 1945. 310 pp.

13. Alarik Rynell, The Rivalry of Scandinavian and Native Synonyms in Middle English, especially *taken* and *nimen*. With an Excursus on *nema* and *taka* in Old Scandinavian. 1948. 431 pp.

14. Henning Hallqvist, Studies in Old English Fractured *ea*. 1948. 167 pp.

15. Gösta Forsström, The Verb 'to be' in Middle English. A Survey of the Forms. 1948. 236 pp.

16. Bertil Widén, Studies on the Dorset Dialect. 1949. 179 pp.

17. Claes Schaar, Critical Studies in the Cynewulf Group. 1949. 337 pp.

18. Bertil Sundby, The Dialect and Provenance of the Middle English Poem *The Owl and the Nightingale*. A Linguistic Study. 1950. 218 pp.

19. Bertil Thuresson, Middle English Occupational Terms. 1950. 285 pp.

20. Karl-Gunnar Lindkvist, Studies on the Local Sense of the Prepositions *In, At, On,* and *To* in Modern English. 1950. 429 pp.

21. Sven Rubin, The Phonology of the Middle English Dialect of Sussex. 1951. 235 pp.

22. Bertil Sundby, Christopher Cooper's English Teacher (1687). 1953. CXVI+10*+123 pp.

23. Björn Wallner, An Exposition of Qui Habitat and Bonum Est in English. 1954. LXXI+122 pp.

24. Rudolf Magnusson, Studies in the Theory of the Parts of Speech. 1954. VIII+120 pp.

25. Claes Schaar, Some Types of Narrative in Chaucer's Poetry. 1954. 293 pp.

26. Börje Holmberg, James Douglas on English Pronunciation c. 1740. 1956. 354 pp.

27. Eilert Ekwall, Etymological Notes on English Place-Names. 1959. 108 pp.

28. Claes Schaar, An Elizabethan Sonnet Problem. Shakespeare's Sonnets, Daniel's Delia, and their Literary Background. 1960. 190 pp.

29. Elis Fridner, An English Fourteenth Century Apocalypse Version with a Prose Commentary. Edited from MS Harley 874 and ten other MSS. 1961. LVIII+290 pp.

30. The Published Writings of Eilert Ekwall. A bibliography compiled by Olof von Feilitzen. 1961. 52 pp.

31. Ulf Jacobsson, Phonological Dialect Constituents in the Vocabulary of Standard English. 1962. 335 pp.

32. Claes Schaar, Elizabethan Sonnet Themes and the Dating of Shakespeare's 'Sonnets'. 1962. 200 pp.

33. Eilert Ekwall, Selected Papers. 1963. 172 pp.

34. Arne Zettersten, Studies in the Dialect and Vocabulary of the Ancrene Riwle. 1965. 331 pp.

35. Gillis Kristensson, A Survey of Middle English Dialects 1290–1350. The Six Northern Counties and Lincolnshire. 1967. XXII+299 pp.

LUND STUDIES IN ENGLISH

(continued from overleaf)

36. Olof Arngart, The Middle English Genesis and Exodus. Re-edited from MS. C.C.C.C. 444 with introduction, notes and glossary. 1968. 277 pp.

37. Arne Zettersten, The English of Tristan da Cunha. 1969. 180 pp.

38. Ellen Alwall, The Religious Trend in Secular Scottish School-Books 1858–1861 and 1873–1882. With a survey of the debate on education in Scotland in the middle and late 19th century. 1970. 177 pp.

39. Claes Schaar, Marino and Crashaw. *Sospetto d'Herode*. A commentary. 1971. 300 pp.

40. Sven Bäckman, This Singular Tale. A Study of *The Vicar of Wakefield* and Its Literary Background. 1971. 281 pp.

41. Christer Påhlsson, The Northumbrian Burr. A Sociolinguistic Study. 1972. 309 pp.

42. Karl-Gustav Ek, The Development of OE *ƿ* and *ðo* in South-Eastern Middle English. 1972. 133 pp.

43. Bo Seltén, The Anglo-Saxon Heritage in Middle English Personal Names. East Anglia 1100–1399. 1972. 187 pp.

44. Kerstin Assarsson-Rizzi, Friar Bacon and Friar Bungay. A Structural and Thematic Analysis of Robert Greene's Play. 1972. 164 pp.

45. Arne Zettersten, A Critical Facsimile Edition of Thomas Batchelor. 1974. 260 pp.

46. Erik Ingvar Thurin, The Universal Autobiography of Ralph Waldo Emerson. 1974. XII+288 pp.

47. Harriet Björk, The Language of Truth. Charlotte Brontë, the Woman Question, and the Novel. 1974. 152 pp.

48. Anders Dallby, The Anatomy of Evil. A Study of John Webster's *The White Devil*. 1974. 236 pp.

49. Gillis Kristensson, John Mirk's *Instructions for Parish Priests*, edited from MS Cotton Claudius A II and six other Manuscripts, with Introduction, Notes and Glossary. 1974. 287 pp.

50. Stig Johansson, Papers in Contrastive Linguistics and Language Testing. 1975. 179 pp.

51. Bengt Ellenberger, The Latin Element in the Vocabulary of the Earlier Makars: Henryson and Dunbar. 1977. 163 pp.

52. Marianne Thormählen, *The Waste Land*. A Fragmentary Wholeness. 1978. 248 pp.

53. Lars Hermerén, On Modality in English. A Study of the Semantics of the Modals. 1978. 195 pp.

54. Sven Bäckman, Tradition Transformed. Studies in the Poetry of Wilfred Owen. 1979. 206 pp.

55. Jan Jönsjö, Studies on Middle English Nicknames. I. Compounds. 1979. 227 pp.

56. A Corpus of English Conversation, edited by Jan Svartvik and Randolph Quirk. 1980. 893 pp.

57. Lars-Håkan Svensson, Silent Art. Rhetorical and Thematic Patterns in Samuel Daniel's *Delia*. 1980. 392 pp.

58. Ingrid Mårdh, Headlinese. On the Grammar of English Front Page Headlines. 1980. 200 pp.

59. Stig Johansson, Plural Attributive Nouns in Present-Day English. 1980. X+136 pp.

60. Claes Schaar, The Full Voic'd Quire Below. Vertical Context Systems in *Paradise Lost*. 1982. 354 pp.

61. Gunilla Florby, The Painful Passage to Virtue: A Study of George Chapman's *The Tragedy of Bussy D'Ambois* and *The Revenge of Bussy D'Ambois*. 1982. 266 pp.

62. Bengt Altenberg, The Genitive v. the *of*-Construction. A Study of Syntactic Variation in 17th Century English. 1982. 320 pp.

63. Jan Svartvik, Mats Eeg-Olofsson, Oscar Forsheden, Bengt Oreström, Cecilia Thavenius, Survey of Spoken English. Report on Research 1975–81. 1982. 112 pp.

Liber

ISBN 91-40-04822-5